Lakshmi S. Iyer
Daniel J. Power   *Editors*

Reshaping Society through Analytics, Collaboration, and Decision Support

Role of Business Intelligence and Social Media

Springer

Annals of Information Systems

Volume 18

Lakshmi S. Iyer • Daniel J. Power
Editors

Reshaping Society through Analytics, Collaboration, and Decision Support

Role of Business Intelligence
and Social Media

Springer

Editors
Lakshmi S. Iyer
Bryan School of Business and Economics
University of North Carolina Greensboro
Greensboro, NC, USA

Daniel J. Power
College of Business Administration
University of Northern Iowa
Cedar Falls, IA, USA

Printed on acid-free paper

Contents

About the Editors

Lakshmi S. Iyer Lakshmi Iyer is an Associate Professor and Director of Graduate Programs in the Information Systems and Supply Chain Management Department, Bryan School of Business and Economics at the University of North Carolina Greensboro (UNCG). She received her Ph.D. in Business Administration from the University of Georgia, Athens, GA and her M.S. in Industrial Engineering from the University of Alabama, Tuscaloosa. Her research interests are in the area of business analytics, knowledge management, emerging technologies and its impact on organizations and users, and diversity in computing. Her research work has been published in or forthcoming in *Communications of the AIS*, *Journal of Association for Information Systems, European Journal of Information Systems, Communications of the ACM, Decision Support Systems, eService Journal, Journal of Electronic Commerce Research, International Journal of Business Intelligence Research, Information Systems Management, Journal of Global Information Technology and Management*, and others. She is a board member of Teradata University Network, Chair of the Special Interest Group in Decision Support and Analytics (formerly SIGDSS), and served as research track co-chair for BI Congress.

Dr. Iyer is also involved in community engaged outreach and scholarship that furthers the role of women in IT (wiit.uncg.edu). She is a member of the American Association of University Women (AAUW) and received the Dr. Shirley Hall Award from AAUW Greensboro Branch in April 2011 for exemplary contribution to enrich STEM education for women. She is founder and Director of the "IT is for Girls" at UNCG, an outreach program for middle and high-school girls that aims to increase their awareness about education and career paths in computing. She has received funding from AAUW, National Center for Women in IT (NCWIT) and from foundations to offer STEM events for young women. Dr. Iyer serves as a co-chair of the

Association of Information Systems' (AIS) task force on Women in IS to enhance the outreach efforts of AIS to women in Information Systems (IS) based on systematic assessment of the current status of women in IS, globally, including students (both current and potential) and professionals in academia, corporate, and non-profit organizations with the intent to creating a nurturing, supporting environment conducive to enhancing the growth and success of women in IS fields.

Daniel J. Power Daniel J. Power is a Professor of Management and Information Systems at the College of Business Administration at the University of Northern Iowa, Cedar Falls, Iowa and the Editor of DSSResources.COM, the Web-based knowledge repository about computerized systems that support decision making, the editor of PlanningSkills.COM, and the editor of DSS News, a bi-weekly e-newsletter. Dr. Power writes a regular column in Decision Support News. Also, Dr. Power is a blogger on the Business Intelligence Network.

Since 1982, Dr. Daniel Power has published more than 50 articles, book chapters and proceedings papers. His articles have appeared in leading journals including Decision Sciences, Decision Support Systems, Journal of Decision Systems, MIS Quarterly, Academy of Management Review, Communications of the Association for Information Systems and Information and Management. He is also co-author of a book titled Strategic Management Skills and he has authored four books on computerized decision support. His DSS Concepts book (2002) titled Decision Support Systems: Concepts and Resources for Managers is a broad ranging scholarly handbook on the fundamentals of building decision support systems. His expanded DSS Framework has received widespread interest. His latest book from Business Expert Press is titled Decision Support, Analytics, and Business Intelligence.

Dr. Power is the Editor-in-Chief of the Journal of the Midwest Association for Information Systems (JMWAIS), a member of two academic journal editorial boards, and was the founding section editor of the ISWorld pages on Decision Support Systems Research and was founding Chair of the Association for Information Systems Special Interest Group on Decision Support and Analytics (SIG DSA). Also, Professor Power was the founding President of the Midwest United States Chapter of the Association for Information Systems (MWAIS) and served as the at-large member of the Board of Directors.

In 1982, Dr. Power received a Ph.D. in Business Administration from the University of Wisconsin-Madison. He was on the faculty at the University of Maryland-College Park from 1982 to 1989. Dr. Power has been a visiting lecturer at universities in China, Denmark, Ireland, Israel, and Russia. Dr. Power has consulted with a number of organizations and in Summer 2003 he was a Visiting Faculty Research Fellow with the U. S. Air Force Research Lab Information Directorate (AFRL/IF).

Dr. Power is a pioneer developer of computerized decision aiding and support systems. During 1975–77, he developed a computerized system called DECAID, DECision AID. In 1981–83, he reprogrammed and expanded the system for the Apple II PC.

Chapter 1
Introduction

Lakshmi S. Iyer and Daniel J. Power

Abstract The Association for Information Systems (AIS) Special Interest Group on Decision Support Systems (SIGDSS) workshop was planned as an event associated with the International Conference on Information Systems (ICIS 2013) in Milan, Italy at Bocconi University from December 14–18, 2013. In keeping with the ICIS2013 theme of "Reshaping Society", the 2013 SIGDSS Workshop aimed to bring together academic and industry professionals from around the world who have a passion for research and education innovation in "Reshaping Society through Analytics, Collaboration, and Decision Support: Role of BI and Social Media". This volume in the Annals of Information Systems reports work originally reviewed for that workshop that were presented and subsequently revised and refined as chapters for this book.

Keywords Introduction • Decision support • Analytics • Workshop • Business intelligence • Data science

In Spring 2013, planning began for the Association for Information Systems (AIS) Special Interest Group on Decision Support Systems (SIGDSS) workshop for December 2013. The workshop was planned as an event associated with the International Conference on Information Systems (ICIS 2013) in Milan, Italy at Bocconi University from December 14–18, 2013.

In keeping with the ICIS2013 theme of "Reshaping Society", the 2013 SIGDSS Workshop aimed to bring together academic and industry professionals from around the world who have a passion for research and education innovation in "Reshaping Society through Analytics, Collaboration, and Decision Support: Role of BI and Social Media".

L.S. Iyer (✉)
Bryan School of Business and Economics, University of North Carolina Greensboro,
435 Bryan Building, UNC-Greensboro, Greensboro, NC 27412, USA
e-mail: Lsiyer@uncg.edu

D.J. Power
College of Business Administration, University of Northern Iowa,
255 Curris Business Building, Cedar Falls, Iowa 50614, USA

This volume in the Annals of Information Systems reports papers originally reviewed for that workshop that were presented and subsequently revised and refined as chapters for this book. One relevant additional chapter that was not presented was also included. Track chairs also summarized presentations, panel comments and other relevant workshop materials not included as chapters.

A major purpose of the workshop was to bring together a core group of leading researchers in the field to discuss where computer involvement in human decisions is headed. Equally important is discussing the role of academic researchers in investigating current and future uses and in creating them. This volume endeavors to find a balance between systematizing what we know so we can teach our findings from prior research better and stimulate excitement to move the field in new directions. Computerized decision support and computing and information technologies can reshape our world, but potentially they can reshape it in negative as well as positive ways.

Teradata University Network (TUN) and Teradata (http://www.teradata.com/) provided financial support for the workshop and helped with recruiting an Industry Keynote speaker. Teradata and TUN sponsored a reception for attendees the evening prior to the workshop.

Teradata University Network is a free, web-based portal that provides teaching and learning tools that is used by more than 45,000 students around the world, majoring in everything from information systems, management or computer science to finance, marketing or accounting. The resources and content support everything from Introduction to IT at the undergraduate level to graduate and executive level courses in big data and analytics. A key to the success of Teradata University Network is that it "is led by academics to ensure the content will meet the needs of today's classrooms".

The workshop began with a Welcome by Ramesh Sharda, a Regents Professor at Oklahoma State University, Executive Director of TUN (teradatauniversitynetwork.com/) and Chair of the SIG DSS Advisory Board for 2012-1013. Then Barbara Dinter, the Panel Chair and Professor at Chemnitz University of Technology, Germany, introduced the Keynote speaker Davide Artioli from Credito Emiliano. Artioli spoke about the business intelligence journey at Credito Emiliano S.p.A. and its impact on reshaping the bank-customer relationship.

During the workshop, seven completed research papers were presented, nine research-in-progress short papers, three teaching and pedagogy papers including the TUN Teaching Award Winner presentation titled "Effective use of Data and Decision Analytics to Improve Order Distribution in a Supply Chain" by Michelle L.F. Cheong and Murphy Choy, Singapore Management University.

Professor Barbara Dinter moderated a distinguished panel titled "A Swiss-Army Knife? The Profile of a Data Scientist." Panelists included: Sudha Ram, Eller College of Management at the University of Arizona; Roger Chiang, Carl H. Lindner College of Business, University of Cincinnati; Detlef Schoder, University of Cologne; and Francesco Mari, Vice President, Global Head SAP HANA Program – SAP Custom Development.

The Research track identified research that promoted theoretical, design science, behavioral research and development of emerging applications in innovative areas

of analytics, collaboration and decision support. The Research track co-chairs were Thilini Ariyachandra, Xavier University, and Amit Deokar, Dakota State University. They coordinated the review and revision process for the workshop and for the research chapters in this volume.

The Research track call for papers noted: Business Intelligence (BI)/ Decision Support (DSS)/ Analytics have become core to many businesses. The social media scape has come to modify and redefine not only businesses but also societal behavior and practices. Although addressed by research in the past few years, these domains are still evolving. For instance, the explosive growth in big data and social media analytics requires examination of the impact of these technologies and applications on business and society. Similarly, traditional approaches to collaboration and decision support are experiencing evolution and growth from BI and social media technologies and trends. As organizations in various sectors formulate IT strategies and investments, it is imperative to understand how various emerging technologies under the BI/DSS umbrella such as big data, mobile and wireless technologies, cloud computing, and recent collaboration tools can be used effectively for decision making in these organizations.

The Teaching track invited contributions that focus on pedagogical practices addressing acquisition, application, and continued development of the knowledge and skills required in the use of analytics, collaboration, and decision support systems. The Teaching track co-chairs were Babita Gupta, California State University Monterey Bay, and Uzma Raja, The University of Alabama, Tuscaloosa.

The Teaching track co-chairs noted in their call for teaching articles and cases studies that "emerging technologies in business intelligence and social media are fueling a need for innovative curricula in online, traditional and hybrid delivery format that meets the industry needs. In keeping with the theme of "Reshaping Society", we are seeking research contributions into the pedagogical practices that address acquisition, application, and continued development of the knowledge and skills required in the usage of analytics, collaboration, and decision support systems with special focus on business intelligence and social media."

1.1 About AIS SIG DSA

By vote of the members in an email referendum completed on Friday, March 7, 2014 the name of this special interest group was changed to the Special Interest Group on Decision Support and Analytics (SIGDSA) – The International Society for Decision Support and Analytics. Revised by-laws have been submitted to the Association for Information Systems (AIS) staff and been approved.

Originally the special interest group was named AIS SIG DSS: ISDSS. The acronym referred to the focus on decision support, knowledge and data management systems (DSS) and to the International Society for Decision Support Systems (ISDSS), an independent predecessor group.

Special Interest Group on Decision Support and Analytics (SIGDSA) Council approved SIG DSS in mid-December 2001. Six AIS members, S. Alter, J. Courtney, R. Sharda, V. Sugumaran, M. Turoff, and H. Watson, joined Daniel Power in submitting the proposal for SIG DSS that was approved. Dan Power served as Founding Chair and Ramesh Sharda was the initial Secretary/Treasurer of SIGDSS.

In Fall 2002, the SIG DSS leadership team was expanded to include Paul Gray (Claremont) as Founding Past Chair, Mike Goul (ASU) as Vice Chair, and Karen "Dowling" Corral (ASU) as Program Chair. At that time the scope was broadened to include knowledge management.

In Fall 2003, the by-laws of AIS SIG DSS were in the drafting process and an agreement was reached to merge the International Society for Decision Support Systems (ISDSS) with SIG DSS. As part of the merger, Clyde Holsapple (U. Kentucky) and Andrew Winston (U. Texas-Austin) joined the SIG DSS Advisory Board and in December 2003 AIS Council accepted the group's by-laws formally creating AIS SIG DSS: ISDSS.

ISDSS was founded in 1989 by Professors Holsapple, Whinston and others "to promote an active flow of ideas and knowledge among persons working in decision support systems and related fields." The International Society for Decision Support Systems (ISDSS) was the first international society devoted to the advancement of decision support system research, practice, and education. In pursuing these goals, ISDSS organized a number of international conferences with refereed papers, plenary speakers and structured idea exchanges. The conferences had registration caps of 80–90 attendees and sought the best of current thinking on DSS topics. The first conference, ISDSS'90, was held at the University of Texas, in Austin, Texas in September 1990. The second ISDSS Conference was held in Ulm, Germany in May 1992 with the theme "Exploring the Next Generation of Decision Support Systems" and it was chaired by Professor Dr. F. J. Radermacher.

At the SIG DSS Business meeting on Sunday, December 15, 2013 following the "Reshaping Society through Analytics, Collaboration, and Decision Support" workshop, the subject of changing the name of the SIG to include analytics was discussed. On February 24, 2014, The SIG DSS Advisory Board approved a name change proposed by the 2013–2014 Chair Laksmi Iyer to Special Interest Group on Decision Support and Analytics (**SIGDSA**) – The International Society for Decision Support and Analytics.

The revised mission of SIG DSA is: To facilitate the exchange, development, communication, and dissemination of information about Decision Support, Analytics, Collaboration and Knowledge Management research and teaching issues in business, management, and organizational contexts among AIS members and in the larger community of IS/IT practitioners and scholars.

An email ballot for the name change proposal was sent out by Secretary/Treasurer Amit Deokar on February 27, 2014, and as noted above current SIG DSS members approved the change.

The 2013 pre-ICIS SIGDSS workshop Co-Chairs were Lakshmi Iyer, University of North Carolina Greensboro, and Daniel J. Power, University of Northern Iowa.

(L-R) Daniel Power, Davide Artioli, Susan Baxley and Lakshmi Iyer. Keynote speaker Davide Artioli is presented with a plaque at the SIGDSS Workshop

Chapter 2
Big Data Panel at SIGDSS Pre-ICIS Conference 2013: A Swiss-Army Knife? The Profile of a Data Scientist

Barbara Dinter, David Douglas, Roger H.L. Chiang, Francesco Mari, Sudha Ram, and Detlef Schoder

Abstract The purpose of the big data panel was to provide a forum for exchange of ideas on curricula content in the area of data science and big data. The panelists were from a broad range of academic institutions designed to provide different perspectives. Industry perspectives are vital as they will the ones employing the graduates of these programs. Thus, the panel included an industry expert from a company that is a leader in data science and big data. Although there was agreement on selected skills as being foundational, it was recognized that a curriculum would not provide all the skills a data scientist would need for many big data projects—thus the team approach to projects.

Keywords Big data • Data science • Data scientist • Curricula • Careers

Panel Coordinators: Barbara Dinter and David Douglas
Panelists: Roger H.L. Chiang, Francesco Mari, Sudha Ram, and Detlef Schoder

B. Dinter (✉)
Chemnitz University of Technology, Thueringer Weg 7, 09126 Chemnitz, Germany
e-mail: barbara.dinter@wirtschaft.tu-chemnitz.de

D. Douglas
Information Systems, University of Arkansas, Business Building 204B,
72701 Fayetteville, AR, USA
e-mail: ddouglas@walton.uark.edu

R.H.L. Chiang
University of Cincinnati, Cincinnati, OH, USA

F. Mari
SAP, Milan, Italy
e-mail: francesco.mari@sap.com

S. Ram
University of Arizona, Tucson, AZ, USA

D. Schoder
University of Cologne, Cologne, Germany
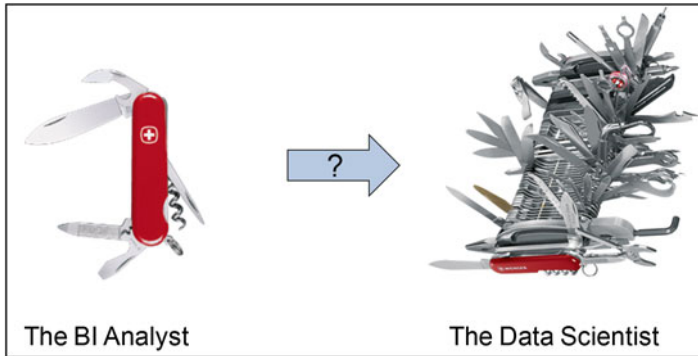
Fig. 2.1 The analogy of the Swiss-army knife (Pictures taken from www.wengerna.com)

Big data as a recent hot topic not only affects the information systems (IS) landscape in industry, it also has significant impact on academia. The need for so-called data scientists has recently attracted widespread attention, in particular initiated by Davenport et al. (2012). Well-known is the Data Science Venn Diagram (Conway 2010) which was a first attempt to illustrate the different skill areas a data scientist should cover. Although differing slightly in detail, there seems to be a consensus that the following skill areas are needed for those profiles (Chen et al. 2012; Chiang et al. 2012; Conway 2010; Davenport et al. 2012; Laney and Kart 2012; Patil 2011): analytical skills, IT and programming skills, business and domain knowledge, and interpersonal skills (such as communication and curiosity).

The panel discussion on "A Swiss-Army Knife? The Profile of a Data Scientist" was aimed at sharing experiences in big data education and identifying best practices for pedagogy to produce well prepared students for a career in industry. The panel title was motivated by drawing an analogy between the capabilities of a data scientist and the functionalities of a Swiss Army knife. While "traditional" analysts need specific but limited skills, such as business intelligence, data warehousing, and data management, the profile of a data scientists seems to be much broader. Based on the analogy, the latter corresponds to the so-called Wenger giant knife as illustrated in Fig. 2.1.

The panelists come from both, academia having a background in big data education and industry, working with data scientists. The panel has been structured by three topic areas and one corresponding "guiding" question. For each of them the panelists were asked in advance to prepare a short statement:

- The profile of a data scientist: What skills does a data scientist need?
- Data science curriculum: Do we need dedicated data science programs? If so, what facilities/disciplines should be in charge of the program?
- Data science career in industry: How students can be motivated and enthused for a data science career?

After a short presentation of the purpose of the panel and an introduction of the panel members each panelist presented his/her point of view for the aforementioned topics. The profile of a data scientist was described similar to the aforementioned data science Venn diagram. Required knowledge and skills focused on several areas including analytical, IT, business, domain, and communications. The analytical knowledge and skills included data mining, statistics, text mining, optimization, sentiment analysis, network and graph analysis, econometrics, and predictive modeling. The IT knowledge and skills included relational database and data warehousing, ETL, OLAP, dashboards and visualization, Hadoop and MapReduce environments, cloud computing, and all types of data management.

Another approach to characterize the required skills of a data scientist was guided by the Hal Varian's definition (Varian 2014) of data science which includes ability to take data and to be able to understand it, to process it, extract value from it, visualize it, and communicate it. Based on this rather process-oriented view specific data science skills can be identified in terms of technologies, techniques, and problem formulation. The technologies include programming languages such as R, Java and Python as well as the Hadoop and MapReduce environments and visualization tools such as D3 and GEPHI. Techniques include mathematical modeling, graph and network theory, machine learning, algorithms, and data structures. Furthermore, data scientists must be able to do problem formulation, have computational thinking and curiosity, ask lots of questions and have the ability to tell a story with the data. Indeed there was quite a bit of consensus from the panelists on the skills needed from a data scientist.

Francesco Mari, representing the practitioner's perspective, noted that only very few data scientists might have such exceptional capabilities. He argued that instead hybrid roles are needed and that it is critical to build teams to cover all required skills.

The panelists emphasized that big data education requires joint efforts and collaboration between academia and industry. The obvious "learning by doing" approach was noted as being more relevant than ever. Francesco Mari raised a point which makes curriculum design even more challenging. "The knowledge of the field is in the data" – i.e., the data scientists along with their domain knowledge are able to have the data tell compelling stories. Even in the same industry and/or the same problem domain, "the data can tell different stories and the corresponding impact of the stories can be huge." The data scientist knowhow of storytelling from the data is more difficult to teach and transfer than traditional IT topics such as programming, etc. Such a variety of cases makes data scientist education even more challenging as not only tool and technology knowledge is required. There is also a need for concrete projects and real data in teaching.

Finally, the panel members presented their assessment of data scientist career paths. Although there is definitely no shortage of careers, often appropriate career paths in organizations do not exist. Careers as data scientists are available in practically every industry including retail, manufacturing, health care, and services. Excellent data scientists skills, e.g. in programming, are not rewarded because their

work is mostly in the background and not visible and thus their value not recognized. Francesco Mari suggested that one potential improvement may be to link them with the R&D department.

After the panelists' presentation, the floor was opened for questions from the audience. The issue around roles of data scientists was raised and the discussion focused on which roles in a data science/big data team make sense. This question is hard to answer, as it always depends from the concrete context and project. Experience from practice shows that in most cases up to three team members can provide all the skills that are required in big data projects. In addition, people with a background in other disciplines (e.g. physics) can enrich such teams.

In conclusion, there is a lot of potential for technical MIS programs to develop inter-disciplinary and innovative curricula in partnership with industry for meeting the burgeoning demand for data scientists.

Biography

Barbara Dinter   is Professor and Chair of Business Information Systems at Chemnitz University of Technology, Germany. She holds a Ph.D. from the Technische Universität München, Germany, where she previously earned a master's degree in computer science. Barbara Dinter worked for several years at University of St. Gallen, Switzerland as a Post-Doc and project manager. In her role as an IT consultant, she worked with a variety of organizations. Her research interests include business intelligence and analytics, big data, data driven innovation, and information management. She has published in renowned journals such as Decision Support Systems, Journal of Database Management, and Journal of Decision Systems, and on conferences such as ICIS, ECIS, and HICSS.

David Douglas   is a University Professor, Co-Director of the Institute for Advanced Data Analytics and Director of Enterprise Systems at the University of Arkansas. He holds a Ph.D. in Industrial Engineering from the University of Arkansas. He teaches a wide variety of information systems subjects with emphasis on enterprise systems and global IT, as well as business intelligence/knowledge management focusing on data mining and data warehouses. He has taught in several countries and presents workshops world-wide. He has received a number of honors and awards including International Educator of the Year (IACIS), Majdi Najm Outstanding Service Award (SAP University Alliances), IBM International Professor of the month and NSF co-principal investigator for Enterprise Computing Community. His research interests include enterprise systems, business intelligence and data analytics. His publications have appeared in various journals including *Communications of the ACM, Decision Sciences Journal of Innovative Education, the Journal of Computer Information Systems, the Journal of Organizational and End User Computing, Information and Management, and the Journal of Management Information Systems,* as well as international, national and regional *Proceedings* of various Conferences.

References

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Chiang, R. H. L., Goes, P., & Stohr, E. A. (2012). Business intelligence and analytics education, and program development: A unique opportunity for the information systems discipline. *ACM Transactions on Management Information Systems, 3*(3), 12:1–12:13.

Conway, D. (2010). The data science Venn diagram. [Online]. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

Davenport, T. H., Barth, P., & Bean, R. (2012). How 'Big Data' is different. *MIT Sloan Management Review, 54*(1), 22–24.

Laney, D., & Kart, L. (2012). Emerging role of the data scientist and the art of data science. Gartner Group. White paper.

Patil, D. (2011). *Building data science teams: The skills, tools, and perspectives behind great data science groups*. Cambridge: O'Reilly Radar.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3–28.

Chapter 3
Creating a Data-Driven Global Society

Daniel J. Power

Abstract Data is captured and analyzed for many purposes including supporting decision making. The expansion of data collection and the increasing use of data-driven decision support is creating a data-driven, global political, economic and social environment. This emerging global society is highly interconnected and many people rely on information technology to support decision making. This chapter explores the impacts that have and might occur as decision support technologies improve and continue shaping global society in new directions. Better understanding of the decision support and analytics phenomenon may help predict future societal changes and consequences. The goal of this analysis is to formulate hypotheses about the impact of decision support for further testing and speculate about long-run consequences. The increasing volume, velocity and variety of data is important to building new decision support functionality. Data collection expansion is part of a self-reinforcing decision support cycle that results in collecting more data, doing more analyses, and providing more and hopefully better decision support. Overall, nine hypotheses are proposed and briefly explored. More research is needed to test and verify them, but anecdotal evidence indicates analytics, business intelligence and decision support are creating a global society that is a data centric, real-time, decision-oriented socio-economic system. Data and decision scientists and technologists should anticipate and ponder the consequences of creating a more pervasive, data-driven global society.

Keywords Data-driven society • Decision support theory • Anticipating consequences • Globalization

D.J. Power (✉)
College of Business Administration, University of Northern Iowa,
255 Curris Business Building, Cedar Falls, Iowa 50614, USA
e-mail: Daniel.Power@uni.edu

3.1   Introduction

Computer-based analysis and decision support began impacting global society in the early 1950s (cf., Power 2008). According to Friedman (1999), globalization has resulted from the democratization of finance, information and technology. Global society refers broadly to an interconnected political, business, economic and social environment. By 1995, the Internet and the World-Wide Web were facilitating wider deployment of analytics and decision support. Smart phones and new data capture technologies like radio frequency identifiers (RFID) have speeded up the changes. Sixty years of decision support technology progress has had many impacts on what managers and decision makers do, how organizations operate, how people think, and what is considered important by people. Organization decision-making is linked to data from many sources and analysis of data is viewed as important. Managers want insight into customer behavior, more predictability in the supply chain and faster, more agile responses in changing, competitive situations. Political and military leaders want more predictability and more agile responses to situations as well. Analytics and decision support are providing solutions for managers in public and private organizations and are creating new capabilities for individuals.

Change related to decision support is definitely occurring in our global society (cf., Power and Phillips-Wren 2011). Some changes like social networks are perhaps minor and seemingly insignificant. Other changes linked to mobile computing and "smart" software seem more disruptive. The forces creating the changes like adoption of predictive analytics are difficult to discretely identify. We live in a multi-causal world of pressures that are tugging, pushing and interacting. One cannot say definitively that the Internet or social networks or even computers are having or have had a profound impact on how we live and work. Perhaps we do live in an emerging hyper-connected, digital, information age. By chance or design we are creating a data-driven, global society and the change process has just begun. Because of information technologies we can certainly interact more easily with people from most countries and often in real-time. Most data is now digital and we are creating and capturing enormous amounts of data.

This article explores the present and future of decision support and analytics and the impacts that have and might occur as decision support technologies change and continue shaping global society. Understanding the decision support phenomenon may help predict future changes and consequences. The goal of this analysis is to formulate hypotheses for further testing and speculate about future consequences. Presently, society is coping with increasingly large and diverse data bases called by some commentators and vendors "big data" that can be and are increasingly used for decision support. Technology progress and the information needs of leaders and managers creates a self-reinforcing cycle of increasing data capture and analysis.

A short essay can only begin to explain what is occurring. The next section explores the large expansion of data available for analysis and decision support and the self-reinforcing decision support cycle, the third section develops hypotheses about what has changed and how society has been reshaped, section four speculates about what might occur and identifies a need to monitor the impacts of decision

support technologies and compile speculative literature related to decision support, the final section summarizes conclusions and raises concerns about some technology paths and notes the need for data and decision scientists to both facilitate change and study the changes that have and are occurring as a result of analytics, business intelligence systems and computerized decision support.

3.2 Data Expansion and Decision Support

There has been a significant change in how much and what types of data organizations capture, retrieve and store. Some call the present the era of "Big Data". Big data is more a term for a marketing phenomenon than a descriptor of data, but it underscores the magnitude of change. Each day, every one of us generates very large amounts of digital data—email, online purchases, using Google Docs, uploading photos to Facebook, using Google Search, and paying bills online. This data and much more from our daily activity is recorded and often backed-up in the Cloud. Most likely someone is or will analyze data from our personal and organization activities.

Organizations store employee created and inputted data, customer and user generated data from Web forms, social media, and even video games, and digital device generated data. Computers and other devices generate digital data. Increasingly there is information technology in everything.

Machine data is a major contributor to the data expansion. Machine data is all of the data generated by a computing machine while it operates. Examples of machine data include: application logs, clickstream data, sensor data and Web access logs (cf., Power 2013b). According to an IBM estimate (2013), each day we create 2.5 quintillion bytes (2.5 Exabytes) of data generated by a variety of sources—from climate information, to posts on social media sites, and from purchase transaction records to healthcare medical images. The estimate is that globally 50,000 Exabytes of data will be generated each year by 2020. The data expansion is increasing exponentially.

Provost and Fawcett (2013) define big data as "datasets that are too large for traditional data-processing systems and that therefore require new technologies" like the Apache™ Hadoop® project that develops open-source software for reliable, scalable, distributed computing (cf., http://hadoop.apache.org/). Ehrenberg (2012) first used the term "big data" in 2009 to label a new ventures fund for "tools for managing large amounts of data and applications for extracting value from that data". That narrow definition morphed into a broader, more amorphous social phenomenon denoting both great potential and concerns about privacy.

Digital data is extremely high volume, it is "big". Data comes from both new and old sources and the increased volume of data led some vendors and industry observers to proclaim the era of 'Big Data'. IBM researchers (Zikopoulos et al. 2013; IBM 2011) describe big data in terms of four dimensions: Volume, Velocity, Variety, and Veracity. Veracity refers to accurate and "true" data and that dimension is the most suspect and controversial. All data captured is not accurate.

Many observers including Aziza (2013), Ehrenberg, Franks (2013), Morris (2012) and Mayer-Schönberger and Cukier (2013) have argued that the potential of using big data to improve our personal lives, help businesses compete, and governments provide services is unbounded. According to Ehrenberg, "Greater access to data and the technologies for managing and analyzing data are changing the world." Somehow big data will lead to better health, better teachers and improved education, and better decision-making. How data is analyzed and presented can change behavior. More data should not however be our goal, rather the goal should be better analysis and decision support using data (cf., Devlin 2013a, b).

Nobel laureate Herbert Simon spent his entire career studying decision making in organizations. He argued in a 1973 article that in a postindustrial society, the central problem will not be how to organize to produce efficiently, but how to organize to make decisions—that is, to process information. Data expansion means more processing of data to create information and a greater need to organize tasks and people to use the information in decision making. Managers in a decision-centered organization must identify, document and track decisions that are made. A data-driven, decision-making organization should encourage and reward use of facts and data by employees. Using facts and data to make decisions in organizations has long been a goal of many managers and academics. The desire for more facts led to the extensive collecting of data.

Decisions are made about many issues in business, non-profit and government organizations using available data. The issue may be operational and hence frequently occurring or strategic and novel and non-routine. Decisions get made with the facts at hand. Even if a manager does not make an explicit decision, ignoring a decision situation or choosing not to decide is a decision. Decision making is an important task in our lives. Ideally decisions should be made more explicitly and with facts, but this ideal is often hard to realize. In a decision-centered organization, managers are and will consciously improve decision processes and will explicitly make decisions. More analytics and decision support enables the creation of decision-centered organizations, and ultimately a decision-centered, data-driven society.

Data is ubiquitous. Data may be streaming to a decision maker or retrieved from a historic data store. Figuring out what data is relevant and what the data means in a decision situation if often however challenging. Data can overwhelm a decision maker and can mislead. Data-driven decision making requires anticipating data and analysis needs and providing the opportunity for people to request analyses of additional data. Analytics involves processes for identifying and communicating patterns, derived conclusions and facts. Raw facts gain additional meaning from analysis.

To be transformative, using data must become part of an organization's decision-making culture. The quest to capture and make available appropriate data and relevant analyses must become an urgent requirement and ongoing priority in the organization. A data-driven organization survives and hopefully prospers based on the quality, provision and availability of data to decision-makers. If data is captured where it is generated and it is appropriately stored and managed for use in decision-making, then analytics and data-driven decision support can become the norm for decision making. Managers are creating data-driven organizations.

Creating a data-driven decision making organization has both technology and human resource challenges. The technology issues continue to evolve and will hopefully be resolved as more data and better, easier to use analytic tools become available. The human resource challenge involves retraining and motivating current employees in analytics and model-driven and data-driven decision support. The human resource challenge is greater than the technology challenge.

For many years managers have been on a journey to capture and analyze data to support decision making. The cycle of data capture, analysis and decision support has produced good results and technology continues to improve and facilitate innovation. Whether intentionally or not, technology forces and managerial desires and needs are creating data-driven organizations and a data-driven society. Figure 3.1 (below) diagrams the self-reinforcing decision support cycle that results in collecting more data, doing more analyses, providing more and better decision support, and hopefully resulting in good outcomes for society.

Figure 3.1 suggests a number of hypotheses that can and should be tested. Data and decision support expansion and proliferation seems inevitable so examining the
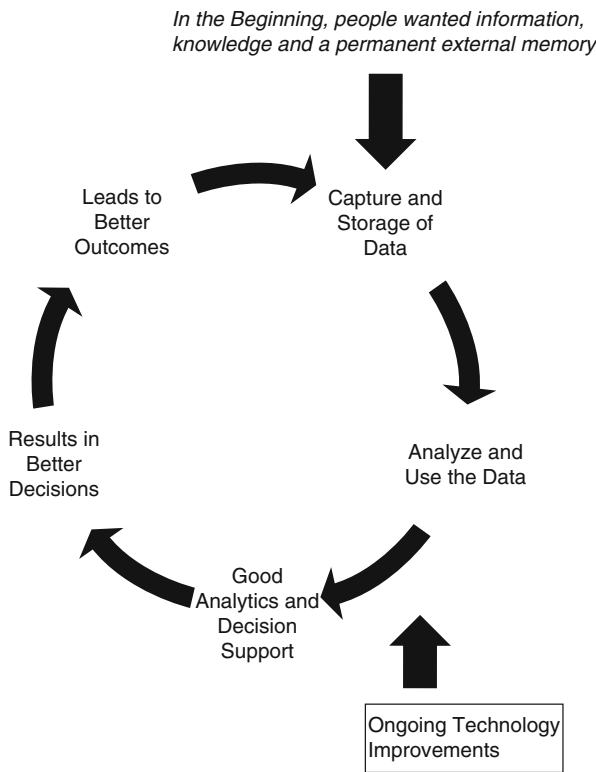
Fig. 3.1 The self-reinforcing decision support cycle

links in the cycle continues to be important. The decision support cycle is anchored in technology innovation and that creates ongoing challenges. For organizations to effectively provide and analyze high volume, high velocity, and high variety data it is necessary to manage and improve: (1) the platforms for storing and accessing data, (2) the analytics, BI and decision support capabilities, and (3) the policies and procedures for governing and managing data including issues of privacy, ethical use and retention (cf., Dyche 2013). Dyche asserts the hard part of big data is managing it. Analyzing big data and making results available in decision situations is also a major challenge. Good analysis and decision support requires data and decision scientists.

The ongoing challenge for decision support and information technology researchers is identifying use cases or user examples for analyzing the large volume of semi- and unstructured data that is accumulating. The decision support cycle continues because managers perceive and anticipate better outcomes as a result of fact-based decision making. So how has analytics and decision support shaped and changed society in developed and developing countries?

3.3   Looking at Today

Modern businesses and many governments have implemented information technologies for decision support. Our global population continues to grow and business and governments encounter increasingly complex decision situations. The rate of adoption varies in organizations, but overall trends can be analyzed. The following are some hypotheses about what has changed because of the decision support cycle and the impact of analytics, BI and decision support on society. Confirming the hypotheses will help validate conclusions in the previous section.

Hypothesis 1 Organizations capture and analyze significantly more data than in the pre-decision support era by many orders of magnitude adjusted for population increases.

McKinsey Global Institute and McKinsey's Business Technology Office studied data volumes in five domains—healthcare in the United States, the public sector in Europe, retail in the United States, and manufacturing and personal-location data globally. The research by Manyika et al. (2011) estimated that, by 2009, "nearly all sectors in the US economy had at least an average of 200 terabytes of stored data (twice the size of U.S. retailer Wal-Mart's data warehouse in 1999) per company with more than 1,000 employees".

According to an Economist Special Report (2010), "There are many reasons for the information explosion. The most obvious one is technology. As the capabilities of digital devices soar and prices plummet, sensors and gadgets are digitising lots of information that was previously unavailable. And many more people have access to far more powerful tools." The article "Data, Data Everywhere" based on an interview with Kenneth Cukier notes "data are becoming the new raw material of business: an economic input almost on a par with capital and labour. "Data capture is increasing for many reasons, but especially because managers hope the data will be useful.

A recent IDC study, reports that researchers in companies and organizations analyze and gather insights from approximately 1 % of the world's data.

Hypothesis 2 Expenditures on decision support technologies are increasing even as the cost per technology unit (per processor, speed of processor, memory, etc.) of capability for hardware falls.

According to Kalakota (2013), Gartner estimates BI and Analytics expenditures grew 16.4 % to USD $12.2 Billion in 2011. A Gartner survey reported at bpmwatch. com "projected that the organizations around the world will spend $2,700 billion in the year 2012 on IT products and services. There is an increase of 3.9 % when compared with 2011. According to Gartner, "Information Technology plays a leading role in the shaping of politics and economy across the world. It has become the leading driver of growth in business."

So computer prices continue to decline and processor capabilities continue to improve. Better computing capabilities should facilitate better computerized analytics, BI and decision support.

Hypothesis 3 Analysis of large quantities of data becomes easier and faster each year even with the increasing volumes of data.

Walker (2013) notes In-Memory Data Grids (IDG) "allow organizations to collect, store, analyze and distribute large, fast-changing data sets in near real-time. … Research shows that organizations using IDG's are able to analyze larger amounts of data at faster speeds than competitors." Contemporary computing capabilities support better and faster data analysis. Remember it doesn't matter if your dataset is "big" or not, what matters is having the capability to analyze the data and turn it the analysis in to actionable recommendations.

Hypothesis 4 Analysts and decision makers can extract and analyze data in real-time and use decision support tools that were not available prior to the advent of wireless and cellular technologies.

Since January 9, 2007, when Apple® introduced the iPhone there have been significant changes in decision support. Decision support on-the-go changes what is possible and what is expected. Managers assume that data and analyses will be available at any time when it is needed.

Hypothesis 5 The total cost of ownership (TCO) of decision support technologies per employee user in constant dollars is lower than at any prior point in the decision support era and the functionality has dramatically increased.

In 2008, a PC had 700 times the capability at the same price as 1977. According to an HP report, "Industry analysts report that only 30 % of IT expenses are hardware-related. The other 70 % are related to costs such as support, maintenance, consumables, end of life and energy." Griliches (2009) notes in a report "A good total cost of ownership (TCO) model incorporates hardware and software costs, installation and license tracking, warranties and maintenance agreements, as well as vendor financing options. It must also include operational expenditures." Many of the costs are fixed costs and evidence suggests the percentage of users in companies is rapidly increasing. Companies are adopting new decision support technologies and incurring more costs. For example, computereconomics.com reported in 2011 "that 28 %

of organizations are currently investing in predictive analytics solutions compared to 22 % that have the technology in place." BUT, the real cost per employee using decision support has declined. Fixed decision support costs are spread over more users and variable costs per user continue to decline.

Hypothesis 6 Much more is now known about customers and citizens. Organizations use the data for data-driven analyses to target activities.

Predictive analytics are increasingly important in large and medium sized organizations. As organizations capture more and more data, it becomes important to analyze and use the data to enhance business results and justify costs. All of the major software vendors market predictive analytics packages and development software packages. For example, IBM advertises "make better decision with business analytics." An IBM web page states "IBM Business Analytics Software takes the guesswork out of business decisions. One company example is MarineMax®, the world's largest boat retailer, that started using IBM Cognos software to inform their inventory decisions. As a result, "their demand planning cycle dropped from 3 months to 3 weeks, leading to a 48 % reduction in costs." Customer Relationship Management (CRM) and predictive analytics have led to extensive collection and analysis of customer data. Managers can identify buying behaviors and patterns using data.

Hypothesis 7 In some situations, information and decision support technologies have reduced information asymmetry.

In multiparty decision situations, one party often has better information than another. These situations commonly involve a purchase/sales transaction or a principal agent situation where a person acts on behalf of another and the principal attempts to monitor and control the agent. Information symmetry means all parties in a decision situation have the same information, an ideal goal. Asymmetric information creates a difficult situation for one or more parties. The problem is harmful when one party has different information and/or better information than another. Deceit, obfuscation and misdirection are all part of the problem. A common example involves selling a car, the owner is likely to have full knowledge about its service history and its likelihood of breaking down. The potential buyer will have less information and may not be able to trust the car salesman. This explains why CARFAX (http://www.carfax.com/) was created and has been successful.

Hypothesis 8 Businesses that use analytics, business intelligence and decision support are more successful and more profitable.

Managers want to believe this hypothesis to justify the large expenditures made on decision support technologies. Establishing a direct causal link between the amount of decision support and profitability is challenging. An indirect indicator is measuring the success of companies where employees are active in decision support vendor users groups like Teradata Partners. One would expect above average returns in these firms that are visible in the decision support community if this hypothesis is true.

Hypothesis 9 Quality of life has improved as a result of analytics, business intelligence and decision support.

This is the proverbial "leap of faith" hypothesis. Many people hope or wish that decision support technologies have improved the quality of life. Proving this hypothesis is difficult. An indirect indicator is whether or not people feel that information and decision support technologies helped them make better personal and organizational decisions.

More research is needed to test and verify these hypotheses, but anecdotal evidence indicates analytics, business intelligence and decision support are shaping global society as a data centric, real-time, decision-oriented socio-economic system. Global society is extremely dependent upon data and data-driven decisions are increasingly the norm in organizations. Business managers want more data and more analyses. The working assumption is that organizations where fact-based, data-driven decision making occurs will have higher profits and better outcomes.

3.4 Looking Forward

Increasingly we live in a crowded world struggling to cope with a growing shortage of physical resources by substituting information technology. The number of 'things' connected to the Internet generating data is increasing dramatically. Cisco says there will be 25 billion things online by 2015; IBM says one trillion. Technology continues to improve and support faster analysis of more and more data in real-time. Technology is enabling change and people are adopting new technologies because of how a specific technology might improve or favorably alter their lives. Also, leaders and managers are relying on technology to understand increasingly complex decision situations.

Change is not directed by a small group of decision makers or controlled. Technologists do not know how a technology like analytics and decision support and its direct application and use in human decisions will alter society. People decide to adopt and use technologies for reasons hidden from careful inspection and review by others. Information technology changes human decision making. Academic researchers have a role in investigating current and future uses of decision support capabilities, but decisions about the implementation of the technologies is decentralized among many people. Researchers do not make choices about whether people or organizations should adopt new decision aiding tools. Those who do make the decisions are primarily considering the impact on their organizations in the next 3–5 years or on themselves in the next year or two rather than considering any long-run consequences for global society.

Decisions about using information technology seem to primarily focus on analyses of short-run financial and operational consequences. Managers and technologists will continue to innovate and adopt new technologies, but researchers need to find a balance between systematizing what we know and speculating about what will result from decision support technology adoption. Decision support researchers should more consciously act to shape the evolving data-driven society rather than accept that social change resulting from our research is a random product of myriad

uncontrolled, interacting events and developments. Perhaps researchers can shape outcomes of technology innovation.

Technology continues to improve and support faster analysis of more and more data in real-time. Also, the computing technology for creating realistic visual simulations has improved tremendously in the past 20 years. For example, Second Life supports an incredible fantasy world where avatars, the representations of the user, can fly, walk through transparent walls, and teleport to imaginary castles, shopping malls and even IBM's robot maze. Avatars and virtual worlds may be part of our decision support future.

Researchers need to better understand the impacts of decision support technologies and stimulate excitement to move the field of decision support in new directions. In 1965, Robert Anthony speculated about the development of large planning data banks. Big data advocates are trying to realize that vision in corporations and government organizations. Computerized decision support, computing and information technologies can improve planning and performance in business organizations, but potentially these technologies will reshape society in negative as well as positive ways. Government and social planners as well as business planners need to use data to support decision-making ethically and appropriately.

Analytics and decision support technologies are changing and reshaping society. Some would say society is being transformed and permanently altered. Perhaps we are creating smarter cities and a smarter planet. The broad question is does decision support make a positive, significant difference in people's lives and perhaps equally important could academics stop the transformation that is occurring if we determined it was important to do so?

To answer these questions it is necessary to engage in speculation about first and second order consequences, to systematize prior research so we can teach our findings, and perhaps to differently organize what we know to change how developers and adopters evaluate decision support options. Helping managers make informed choices about decision support is part of the responsibility of researchers and developers in this field of study.

Technology speculation has been occurring for many years, but future fiction literature is considered entertainment rather than content that might contribute to our understanding of both intended and unintended consequences of adopting decision support or other technologies. Stories by authors like Isaac Asimov and Robert Heinlein are both entertaining and useful to researchers and technologists.

In a story by Heinlein (1964) titled "Beyond this Horizon" that was initially published in Astounding Science Fiction magazine in April 1942-May 1942, we read of a future earth where science and technology have transformed and reshaped society. A computer system with data on all of the financial transactions of a continent predicts economic activity and recommends changes in "the subsidy on retail transfers of consumption goods" and changes in "the monthly citizen's allowance" to maintain a stable social equilibrium (cf., p. 7). The computing system supports decision making, but its predictions are accepted as true and political decision makers have no real choice but to accept the "recommendations".

Isaac Asimov, a prolific science and science fiction author, wrote the Foundation trilogy in the 1950s. The story revolves around the consequences of developing a

decision support model that predicts the future. Mathematician Hari Seldon proposes and then creates a psychological model and a science called psychohistory. The computer model combines data and equations about history, sociology, psychology, and mathematical statistics to make general predictions about the future behavior of very large groups of people. Seldon explores many alternative futures that span thousands of years and tests various interventions that change outcomes. It turns out as the story develops that his predictions are uncannily accurate until chance and human mutation alter the course of predicted events. In the short run, a few hundred years, the decision support forecasting model is accurate, but ultimately people must respond and act more independently to survive.

In Mack Reynold's novela "Computer War" (1967), the world is divided into two nation states. Only one, Alphaland, has computers. The computer predicts victory, but the war goes on without explanation. For some reason the computer's conclusion of Alphaland's economic superiority over Betastan does not lead to a single world state. Computer support and rational analysis was not able to help the leaders of Alphaland defeat a seemingly irrational and unpredictable human foe.

British author John Brunner wrote "Stand on Zanzibar" (1968) about an over-populated dystopian world of 2010. A major plot element is a supercomputer named Shalmaneser that can listen and scan conversations for key words. In the novel there is discussion of whether Shalmaneser is self-aware. It seems unlikely that a super-computer like Shalmaneser can solve humanity's future problems.

Perhaps the best known supercomputer in science fiction is HAL 9000. HAL is a star of Arthur Clarke's (1968) novel and the associated movie directed by Stanley Kubrick titled "2001: A Space Odyssey" (1968). The movie and novel deal with many issues including humanity's move beyond computer decision support to artificial intelligence and its possible consequences. HAL 9000 has some type of breakdown or malfunction on a space voyage and acts to defend itself. Ultimately HAL kills some crew members and so the mission pilot and scientist Dr. David Bowman disconnects Hal's processor core. A thinking machine intended to help people inexplicably harms them and must be disabled.

There are many scenarios that are worth considering. Sun Microsystems cofounder Bill Joy's (2004) Wired article, "Why the Future Doesn't Need Us" warned of the catastrophic potential for twenty-first century technologies like robotics, genetic engineering, and nanotech. Subsequently, he called for technological relinquishment or giving up new technology. Perhaps the future is a Cybertron-like Earth (ala the Transformers), or a society watched over by benevolent machines, or a future where humans and computational machines work together to make the world a better place.

Many more novels, novellas and short stories speculate on the impact of computerized decision support and artificial intelligence on human society. After reading many such stories, it is challenging to recall a decision support innovation that led to a uniformly positive set of consequences and outcomes for humanity. Managers and technologists seem faced with difficult choices of development, adoption and implementation. Compiling scenarios about alternative decision support futures can potentially assist business and political decision makers as they grapple with funding research and purchasing systems to assist in military,

government and business decision situations. The future is unknown, but it is worthwhile contemplating what might happen in the long-term if various decision support and analytic technologies improve substantially and are adopted.

3.5 Conclusions and Commentary

Many years of improvements in computing technology has led to faster computation, faster transmission of data and larger storage capacity. The information age is still at its beginning, but decision automation, technology adoption and decision support applications have increased dramatically. Will smart machines run the world? Will people be relegated to menial, make work tasks? Will decision automation change the world? Today in some situations, computer software does make better decisions than human decision makers. Our future reality is unlikely to be The Matrix (1999) where sentient machines run the world and most people are comatose directly connected to a virtual world, but the development of thinking machines like Hal 9000 is increasingly likely.

Results from a poll at thinkartificial.org (2007) suggest some serious concerns about decision technologies. The respondents were primarily Digg (http://digg.com/) readers. Readers were asked: Do you, for some reason, fear the current and/or future increase of artificial intelligence? 1,002 respondents (16.7 %) checked Yes, I find the idea of intelligent machines frightening. 27.1 % indicated No, I don't find intelligent machines frightening (1,632 votes). Finally, 3366 respondents (56.3 %), indicated I'm not afraid of intelligent machines, I'm afraid of how humans will use the technology. The perceived threats to privacy are real and should be an ongoing concern. Data collected for one purpose can be used for other less noble purposes, and data analysts and data scientists do mistakenly interpret data sets.

Some academics have been skeptical of the vision of creating comprehensive databases for planning and strategic decision making. For example, Harvard Professor Robert Anthony (1965) argued "It is because of the varied and unpredictable nature of data required for strategic planning that an attempt to design an all-purpose internal information system is probably hopeless. For the same reason, the dream of some computer specialists of a gigantic bank, from which planners can obtain all the information they wish by pressing some buttons, is probably no more than a dream (p. 45)." Perhaps Anthony and others are wrong and we can expect much more powerful data-driven decision support with gigantic data banks to assist with strategic planning and operations management. Analysts like Nigel Rayner seem to think so.

Rayner (2011), a Gartner analyst, speculated that "In the next 40 years analytics systems will replace much of what the knowledge worker does today. In other words, systems like IBM's Watson will be your boss and humans—especially the species known as middle management—will go extinct." Rayner argued many tasks middle managers do today will be automated.

According to Rayner, "We are at a tipping point in the evolution of the 'Information Age,' but business culture is holding back the use of IT. In the future,

decision making will be automated and managed by machine-based models far better than any human could manage." If this transformation happens, there will be many changes to business, society and the economy.

Thinking machines are part of the hope for dealing with complexity. Watson (IBM ForwardView 2012), a natural language processing system from IBM, demonstrated in 2011 on a TV game show that machine processing of English phrases has great potential. Watson searched through millions of documents and returned a correct answer in a few seconds. So perhaps middle managers have cause for concern.

Are we at the tipping point? The Kurzweil (2005) singularity? Perhaps. The technological singularity is when artificial intelligence will exceed human intelligence, radically changing civilization. Faster computation, faster transmission of data, and larger storage capacity is a reality and the improvements are continuing. Managers can obtain more data about more topics in real-time. But the desire for more decision support does not necessitate decision automation and thinking machines. The path can lead to improved human decision making supported by technology.

There is a problem associated with making decisions. Evidence indicates many people are poor decision makers much or most of the time. People make bad choices. The goal of decision support has always been to help people make better decisions rather than automating decision-making. Currently, some pundits hope that training more data and decision scientists will help business and government leaders cope with the increasing complexity of decision making in a data rich environment. The new term data science refers to a more sophisticated and systematic analysis of data by experts (Davenport and Patil 2012). Data Scientists will supposedly use "Big Data" and create a context and story that is useful 500 (cf., Power, 2013a,c,d).

The role of data and decision scientists and technologists should include anticipating and thinking about the consequences of creating a more pervasive, data-driven global society. What can now be concluded about how analytics and decision support has and is reshaping society?

First, global society is extremely dependent upon data and data-driven decisions are increasingly the norm in organizations.

Second, technology forces and managerial desires and needs are creating a data-driven society. Data and decision scientists will create new information technology capabilities to analyze the data of the future.

Third, the decision support cycle continues because managers perceive and anticipate better outcomes from more data and more decision support.

Fourth, the ongoing challenge for decision support and information technology researchers is identifying use cases or user examples for analyzing the large volume of semi- and unstructured data.

Fifth, more research is needed to test and verify the nine hypotheses developed about changes that have occurred due to information technology and decision support.

Sixth, managers and technologists face difficult to development, adoption and implementation of innovative decision support.

Seventh, Government and business planners need to use data to support decision-making ethically and appropriately. Compiling scenarios about alternative decision support futures can potentially assist them.

Eighth, data and decision scientists need to help decision makers develop policies about privacy, data retention, security, and data access.

Ninth, we all need to think about the consequences of more and better analytics and decision support. Thinking about the future is useful.

The experiences and developments of past decades show that technology progress has its own often unpredictable timeline and yet the speed of technology and decision support innovation seems to be increasing exponentially. Trends and speculation suggest we humans have many reasons to be hopeful and fearful about the prospects of developing more sophisticated analytics, decision support and thinking machines.

Is a data-driven global society desirable? Technology progress is not a smooth path. Spinoza noted in 1677 "There is no Hope without Fear, and no Fear without Hope."

Biography

Daniel J. Power   Professor of Management and Information Systems at the College of Business Administration, University of Northern Iowa, Cedar Falls, Iowa and the editor of DSSResources.COM, and Decision Support News (aka DSS news), a bi-weekly e-newsletter. Since 1982, he has published more than 40 articles, book chapters and proceedings papers related to decision support. His DSS Concepts book (2002) titled Decision Support Systems: Concepts and Resources for Managers is a broad ranging scholarly handbook on the fundamentals of building decision support systems. His latest book from Business Expert Press is titled Decision Support, Analytics, and Business Intelligence.

References

Anthony, R. N. (1965). *Planning and control systems: A framework for analysis*. Cambridge, MA: Graduate School of Business Administration, Harvard University.

Aziza, B. Big Data 'A-Ha' Moment? *Forbes CIO Central*, February 25, 2013 at URL http://www.forbes.com/sites/ciocentral/2013/02/25/big-data-a-hamoment

Bruner, J. (1968). *Stand on Zanzibar*. New York: Ballantine Books.

Clarke, A. (1968). *2001: A space odyssey*. New York: Signet.

Davenport, T. H., & Patil, D. J. (2012, October). Data scientist: The sexiest job of the 21st century, *Harvard Business Review*.

Devlin, B. (2013a, February 5). Big analytics rather than big data. *B-eye-Network blog*. At URL http://www.b-eye-network.com/blogs/devlin/archives/2013/02/big_analytics_r.php

Devlin, B. (2013b, March 4). Big data—Please, drive a stake through its heart!. *B-eye-Network blog*. At URL http://www.b-eye-network.com/blogs/devlin/archives/2013/03/big_data_-_plea.php

Dyche, J. (2013, March 13). Big data's three-legged stool. *Information Management*. At URL http://www.information-management.com/news/big-data-three-legged-stool-10024077-1.html

Economist Special Report. (2010, February 25). Data, data everywhere. *Economist*. At URL http://www.economist.com/node/15557443

Ehrenberg, R. (2012, January 19). What's the big deal about Big Data? *InformationArbitrage.com blog post*. At URL http://informationarbitrage.com/post/16121669634/whats-the-big-deal-about-big-data

IBM ForwardView (2012, April). Watson's next conquest: Business analytics. At URL www-304.ibm.com/businesscenter/cpe/html0/230318.html?subkey=subscribe&ca=fv1204&me=feature 1&re=ushometxt

Franks, B. (2013). *Taming the big data tidal wave*. Hoboken: Wiley.

Friedman, T. (1999). *The Lexus and the olive tree: ?Understanding globalization*. New York: Farrar Straus Giroux.

Griliches, E. (2009, September). The impact of a total cost of ownership model. *Cisco*. At URL www.cisco.com/en/US/prod/collateral/routers/ps9853/impact_of_total_cost_ownership_model__idc_0928.pdf

Heinlein, R. A. (1964). *Beyond this horizon*. New York: Signet Books.

IBM. (2013). What is big data? At URL http://www-01.ibm.com/software/data/bigdata/. Accessed 6 Mar 2013.

Joy, B. (2004). Why the Future Doesn't Need Us. Wired. At URL http://archive.wired.com/wired/archive/8.04/joy_pr.html

Kalakota, R. (2013, April 24). Gartner says—BI and analytics a $12.2 Bln market. At URL http://practicalanalytics.wordpress.com/2011/04/24/gartner-says-bi-and-analytics-a-10-5-bln-market/

Kurzweil, R. (2005). *The singularity is near*. New York: Viking Books.

Manyika, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011, May). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Morris, J. (2012, July 16). Top 10 categories for Big Data sources and mining technologies. *ZDNet*. At URL http://www.zdnet.com/top-10-categories-for-big-data-sources-and-mining-technologies-7000000926/

Power, D. J. (2008). Decision support systems: An historical overview. In F. Burstein & C. W. Holsapple (Eds.), *Handbook on decision support systems* (Vol. 1, pp. 121–140). Berlin: Springer.

Power, D. J. (2012, April 1). Will thinking machines make better decisions than people? *Decision Support News*, Vol. 13, No. 7.

Power, D. J. (2013a). *Decision support, analytics, and business intelligence* (2nd ed.). New York: Business Expert Press.

Power, D. (2013b, January). What is machine data? *Decision Support News*, Vol. 14, No. 02.

Power, D. J. (2013c, March 17). Does the term big data have utility for managers? *Decision Support News*, Vol. 14, No. 06.

Power, D. J. (2013d, June 23). What is a data scientist. *Decision Support News*, Vol. 14, No. 13.

Power, D. J. (2013e, August 4). How will decision support technologies shape our global society? *Decision Support News*, Vol. 14, No. 16.

Power, D. J. (2013f, September 1). How is analytics, BI and decision support shaping global society? *Decision Support News*, Vol. 14, No. 18.

Power, D. J., & Phillips-Wren, G. (2011). Impact of social media and Web 2.0 on decision-making. *Journal of Decision Systems, 20*(3), 249–261.

Provost, F., & Fawcett, T. (2013). *Data science for business: Fundamental principles of data mining and data-analytic thinking*. Sebastopol: O'Reilly.

Rayner, N. (2011, October 7). Maverick research: Judgment day, or why we should let machines automate decision making. *Gartner*. At URL http://www.gartner.com/id=1818018

Reynolds, M. (1967). *Computer war*. New York: Ace Books.

The Matrix. Warner Bros. Pictures, 1999 at URL http://www.imdb.com/title/tt0133093/

Thinkartificial.org, Results from a poll April 22, 2007 at http://www.thinkartificial.org/web/the-fear-of-intelligent-machines-survey-results/

Walker, M. (2013, March 13). In-memory data grids allow data science at faster speeds. At URL
    www.datasciencecentral.com/profiles/blogs/in-memory-data-grids-allows-data-scientists-
    analyze-big-data-at
Zikopoulos, P., DeRoos, D., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. (2013).
    Harness the power of big data: The IBM big data platform. New York: McGraw Hill.

Chapter 4
Agile Supply Chain Decision Support System

Jaehun Lee, Hyunbo Cho, and Yong Seog Kim

Abstract Recently, many organizations intend to make their supply chains more responsive to the change in demand in terms of volume and variety and hence consider agility one of the most critical evaluation criteria in addition to other well-known criteria such as general management capability, manufacturing capability, and collaboration capability. This paper formulates the supplier evaluation and selection problem as a multi-criteria decision-making (MCDM) problem with subjective and fuzzy preferences of decision makers over available evaluation criteria and provides the decision maker with a decision support system that presents the Pareto fronts, a set of best possible high-quality suppliers and optimized business operation levels from such suppliers. In addition, this paper quantifies the importance of agility and its sub-criteria in the process of evaluating and selecting agile suppliers by measuring the magnitude of bullwhip effect as a measurement of the business impact of resulting agile supply chain. The proposed system based on fuzzy analytic hierarchy process (AHP) and fuzzy technique for order performance by similarity to ideal solution (TOPSIS) successfully determines the priority weights of multiple criteria and selects the best fitting supplier after taking the vagueness and imprecision of human assessments. More importantly, it presents approximated Pareto fronts of resulting supplier chains as the priority weights of agility criterion and sub-criteria within agility are varied.

Keywords Supplier selection • Agile supply chain • Pareto fronts, Bullwhip effect • Fuzzy AHP • Fuzzy TOPSIS

J. Lee • H. Cho
Department of Industrial and Management Engineering, Pohang University of Science
& Technology, San 31 Hyoja, Pohang, Gyungbuk, Republic of Korea
e-mail: jaehun_lee@postech.ac.kr; hcho@postech.ac.kr

Y.S. Kim (✉)
MIS Department, Jon M. Huntsman School of Business,
Utah State University, Logan, UT 84322-3515, USA
e-mail: yong.kim@usu.edu

4.1   Introduction

Due to increasing reliance on outsourcing of many complex services and products, evaluating and selecting qualified suppliers with the highest potential to meet buyers' requirements becomes an essential part of building successful supply chain management (SCM) and sustaining success in business operations (Araz et al. 2007). Often, the supplier evaluation and selection problem is formulated as a multi-criteria decision-making (MCDM) problem with various quantitative and qualitative evaluation criteria such as sales price, market performance, manufacturing capability and delivery time. In this framework, each supplier s_i is associated with an evaluation vector $F = F_1(s_i),\ldots, F_n(s_i)$ where n is the number of quality criteria. One supplier s_1 is said to *dominate* another supplier s_2 if $\forall c: F_C(s_1) \geq F_C(s_2)$ and $\exists c:$ $F_C(s_1) > F_C(s_2)$, where F_C is the evaluation score of the *c-th* criterion, $c \in \{1,\ldots,n\}$. Neither supplier dominates the other if $\exists c_1,c_2: F_{C1}(s_1) > F_{C1}(s_2), F_{C2}(s_2) > F_{C2}(s_1)$. We define the *Pareto front* as the set of non-dominated suppliers or optimized business operation levels from such suppliers. In supplier selection as a Pareto optimization, the goal is to approximate as best possible the Pareto front, presenting the decision maker with a set of high-quality suppliers and optimized business operation levels from which to choose.

Various methods to select the best-fit suppliers while minimizing an organization's spending of money and time have been proposed from diverse disciplines including management, operations research, artificial intelligence, and decision analysis. In particular, turbulent and volatile markets due to the shortened life-cycles and global competitions enforce organizations with lengthy and slow-moving supply chains to review and restructure their supply chains to maximize their business success by meeting the peaks and troughs of customer demands for ever-shorter delivery times (Christopher 2000). To this end, organizations are required to make their supply chains more responsive to the needs of the market through a high level of maneuverability. Naturally, a new evaluation criterion, agility, becomes one of the most critical evaluation criteria for supplier evaluation and selection.

Agility is defined as the ability of an organization to respond rapidly to change in demand, both in terms of volume and variety (Christopher and Towill 2001). Thus, the responsiveness in volume and variety are the key to measuring agility of a supplier. Note that change in demand can come from several different sources such as marketplace, competition, customer desire, technology and social factors (Lin et al. 2006). Originating from the concept of flexible manufacturing systems (FMS) that enables rapid change and a greater responsiveness to changes in product mix or volume through automation, agility in these days is regarded as a business-wide capability that embraces organizational structures, information systems, logistics processes and mindsets (Christopher 2000). Therefore, the resources required for agility are often difficult to mobilize and retain by single companies in the current networked business environments for efficient and effective collaboration across organizations; Hence it is imperative for organizations to cooperate and leverage

complementary competencies along up- and down-stream of their supply chains (Yusuf 2004). In the end, building an agile supply chain through selecting agile suppliers is key to business success, enabled by more responsive business operations, and sustain and strengthen market competitiveness (Kumar et al. 2004).

In this paper, we intend to quantify the importance of agility in the process of evaluating and selecting agile suppliers, and quantify the business impact of resulting agile supply chain on an organization's business operation by measuring the bullwhip effect. Ultimately, we like to present a decision support system that considers not only multiple evaluation criteria for supplier evaluation and selection but also subjective preferences of decision makers over available evaluation criteria. To this end, we first comprehensively review numerous criteria that have been used to evaluate candidate suppliers and classify them into four main domains such as general management capability, manufacturing capability, collaboration capability, and agility. After we identify four main domains of criteria and sub-criteria in each domain, we structure the decision hierarchy by determining the relative importance of each main domain and sub-criteria within the same main domain. Note that assessing the relative importance of decision criteria involves a high degree of subjective judgments, and individual preferences and linguistic assessments based on human feelings and judgments are at most vague. Therefore, we calculate prior weights of decision criteria using fuzzy analytic hierarchy process (AHP) method by presenting linguistic assessments of decision makers on the relative importance of decision criteria not as exact numbers but as interval judgments (Chen et al. 2006). Then we determine the rankings of suppliers using fuzzy technique for order performance by similarity to ideal solution (TOPSIS) and select the best fitting supplier to complete an imaginary supply chain. By successfully integrating AHP and TOPSIS based on the fuzzy theory, we take care of the vagueness and imprecision of human assessments but also reflect the subjective preferences of decision makers, making the presented model generalizable to the cases of decision makers with different preferences.

More importantly, we quantify the importance of agile criterion by comparing the differences of business operations in two supply chains: agile supply chain with the chosen supplier after considering all four main domains including agility criterion and non-agile supply chain with the chosen supplier without considering agility. The business operations of two supply chains are estimated by measuring the magnitude of bullwhip effects assuming that these simple supply chains consist of one buyer and one supplier and demand forecast follows first-order autoregressive model, AR(1). In particular, we approximate the Pareto fronts of agile and non-agile supplier chain as we vary the relative importance of agility criterion in the evaluation and selection process of candidate suppliers. Technically, the Pareto front of agile supply chain in our study represents the magnitudes of bullwhip effects with a set of best suppliers for varying relative importance of agility criterion. The managerial and practical implication of our findings is then to present the decision maker with a set of high-quality suppliers from which to choose as their strategic preferences over agility criterion change. In addition, we estimate the impact of sub-criteria in agility domain on the magnitude of bullwhip effects and draw managerial insights.

The remainder of this paper is organized as follows. In the following section, we provide a literature review on agile supply chain, supplier evaluation and selection, and explains the underlying methodologies used in this study. Then we describe the framework of the proposed agile supply chain DSS for supplier evaluation and selection. Business impacts of agile and non-agile supply chains are assessed by measuring the magnitude of bullwhip effects in the following section. Finally, we provide concluding remarks and suggest future research direction.

4.2 Literature Review

4.2.1 Supplier Selection for Agile SCM Via fuzzy MCDM

This study takes a holistic approach to leverage three closely related domains toward successful agile supply chain construction. The first relevant domain is agile SCM discipline that recognizes the importance of agile supply chain based on the observation that it is supply chains (not companies) that compete against each other to respond rapidly to change in demand both in terms of volume and variety originated from various agility drivers such as marketplace, competition, customer desire, technology and social factors (Lin et al. 2006). In addition, the resources required for agility are often difficult to mobilize and retain by single companies; therefore it is imperative for companies to cooperate and leverage complementary competencies (Yusuf 2004). To this regard, many studies acknowledge that agility is a business-wide capability that embraces organizational structures, information systems, logistics processes, and mindsets (Power et al. 2001), and agility should reflect the ability of an organization to respond rapidly to unexpected demand fluctuations in terms of volume and variety, which can be estimated by measuring the degree of information technology and process integration among suppliers (Christopher and Towill 2001; Krishnamurthy and Yauch 2007). Based on comprehensive literature review and interviews with industry experts, we compile and present a new set of agility evaluation criteria in Table 4.2. General requirements for preferred suppliers (not necessarily agile suppliers) are also compiled and listed in Table 4.2.

The second associated domain is multi-criteria decision making (MCDM) domain that has presented numerous algorithms and models for MCDM problems in marketing, decision science, operation management, computer science, and engineering disciplines (Kim et al. 2005; Wallenius et al. 2008). In this study, we are particularly interested in various methods for supplier evaluation and selection including individual methods (e.g., DEA, AHP, ANP, GP, and ANN) and integrated methods of various algorithms (e.g., AHP-GP). Readers who are interested in comprehensive review of such methods and studies are strongly advised to refer to Ho et al. (2010). Several other studies take another step to integrate these methods with fuzzy approaches to consider the ambiguity and subjectivity of decision making (Bottani and Rizzi 2008; Chan et al. 2008; Chen et al. 2006; Önüt et al. 2009). However, most of these studies are limited in the sense that they are lack

of quantitative analyses of estimating positive business impacts and drawing managerial insights to help decision makers configure a best fitting supply chain to their organizations. The third domain considered in this study is mathematical or organizational science models that provide theoretical models of supply chains and estimate the sensitivity of internal and external factors of supply chains (Bray and Mendelson 2012; Lee et al. 1997, 2000; Klein and Rai 2009; Yao and Zhu 2012). However, these models are limited in the sense that no practical decision making process and subjective assessments of business impacts based on fuzzy multi-criteria are explicitly considered. By leveraging pros and cons of these three domains, this study presents a new decision support system designed toward optimal agile supply chain that reflects fuzzy MCDM process, and estimates its business impacts in terms of the magnitude of bullwhip effects and presents Pareto fronts that help decision makers draw managerial insights.

4.2.2   Fuzzy AHP

In this paper, fuzzy AHP is used to determine the relative weights of evaluation criteria. Therefore, we first briefly review AHP method (Saaty 1980) introduced to solve various qualitative and quantitative MCDM problems involving subjective judgments such as determining the importance of factors in product design and selecting the best product concept (Raharjo et al. 2008). Technically, AHP determines the relative priority (w_i) of the *ith* criterion by exploiting a multi-level hierarchical structure of decision-making problems. Given a set of decision criteria $\{1,\ldots,n\}$, the AHP method starts to construct a pairwise comparison matrix $\mathbf{A}(n \times n)$ whose element a_{ij} ($\neq 0$) represents the relative importance of the *ith* criterion compared with the *jth* criterion by using pre-defined numerical comparison scale scores $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, 1 for equally preferred and 9 for absolutely preferred, by carrying out full pairwise comparisons of n decision criteria. By definition, the value of a_{ii} d a_{ji} set to 1 and $1/a_{ij}$, respectively. The relative normalized weight of the *ith* decision criterion is then determined by calculating the geometric mean of the *ith* row of a pairwise comparison matrix $\mathbf{A}$. The final weights for each criterion are calculated from this final comparison matrix.

However, the requirement of obtaining full pairwise comparisons of evaluation criteria based on pre-defined numerical comparison scale scores of AHP (i.e., $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$) cannot suitably address the uncertainties and ambiguities involved in the assessment of evaluation criteria with linguistic variables more frequently adopted by decision makers. That is, often decision makers prefer using linguistic variables such as "(the *ith* criterion is) Equally (preferred to the *jth* criterion), Weakly, Fairly Strongly, Very Strongly, Absolutely" to using numerical comparison scale scores in the pairwise comparison process. Therefore, it is necessary to transform evaluation inputs based on linguistic variables into numerical numbers for each criterion while dealing with reasoning that is approximate rather than fixed and exact by allowing partial truth, where the truth-value ranges between 0 and 1.

Fig. 4.1 Linguistic variables for the importance weight of each criterion
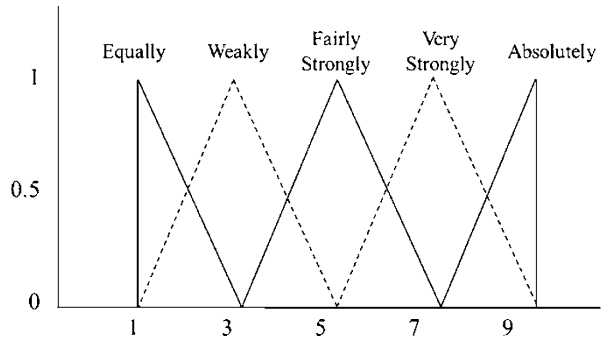
Table 4.1 Linguistic variables and triangular membership functions

Linguistic scale of importance	Fuzzy numbers for the FAHP	Membership function	Domain	Triangular fuzzy number (l, m, u)
Equally preferred	$\tilde{1}$	$\mu_M(x) = (3-x)/(3-1)$	$1 \leq x \leq 3$	$(1.0, 1.0, 3.0)$
Weakly preferred	$\tilde{3}$	$\mu_M(x) = (x-1)/(3-1)$	$1 \leq x \leq 3$	$(1.0, 3.0, 5.0)$
		$\mu_M(x) = (5-x)/(5-3)$	$3 \leq x \leq 5$	
Fairly strongly preferred	$\tilde{5}$	$\mu_M(x) - (x-3)/(5-3)$	$3 \leq x \leq 5$	$(3.0, 5.0, 7.0)$
		$\mu_M(x) = (7-x)/(7-5)$	$5 \leq x \leq 7$	
Very strongly preferred	$\tilde{7}$	$\mu_M(x) = (x-5)/(7-5)$	$5 \leq x \leq 7$	$(5.0, 7.0, 9.0)$
		$\mu_M(x) = (9-x)/(9-7)$	$7 \leq x \leq 9$	
Absolutely preferred	$\tilde{9}$	$\mu_M(x) - (x-7)/(9-7)$	$7 \leq x \leq 9$	$(7.0, 9.0, 9.0)$

To this end, we first define a set of fuzzy numbers (i.e. $\tilde{1}$ $\tilde{3}$, $\tilde{5}$ $\tilde{7}$ $\tilde{9}$ for each criterion (a tilde "~" symbol is placed above to indicate its fuzzy property), and define a fuzzy triangular relationship between each linguistic variable and one of these fuzzy numbers (e.g., "Equally" to $\tilde{1}$ In this study, the linguistic variables are denoted as positive triangular fuzzy numbers for each criterion as shown in Fig. 4.1.

However, due to the fuzzy relationship, each linguistic variable denotes a range of fuzzy numbers with partial membership grade. Then a triangular membership function is defined to define a partial membership grade by a value between 0 and 1, and a resulting triangular fuzzy number is denoted as (*l, m, u*) where the *l, m,* and *u* denote the smallest possible value, the most promising value, and the largest possible value (Chan et al. 2008). Note that when $l = m = u$, it becomes a non-fuzzy number by convention. Namely, the value of fuzzy set means the membership grade of the element x in a fuzzy set. For example, the value of "Equally" fuzzy set for $x = 1$ is 1, and the value for $x = 2$ is 0.5, while the value of "Weakly" fuzzy set for $x = 2$ is 0.5, and the value for $x = 3$ is 1. Then, each linguistic variable may represent numbers with the partial membership grade which is resulted from membership function. We present a triangular membership function and triangular fuzzy scale for each linguistic variable in Table 4.1.

4.2.3 Fuzzy TOPSIS

Once the relative weights of each evaluation criterion of candidate suppliers determined by FAHP, fuzzy TOPSIS is then used to rank candidate suppliers. The TOPSIS (Hwang and Yoon 1981) is based on the concept that the chosen supplier should have the shortest geometric distance from the positive ideal solution and the longest geometric distance from the negative ideal solution. Note that the positive ideal solution consists of all of best values attainable of criteria, whereas the negative ideal solution is composed of all worst values attainable of criteria. To this end, decision makers provide their valuable inputs in regard to how ideal each candidate supplier is for each criterion using a set of numerical scale scores. Once a normalized numerical value for each criterion over a set of candidate suppliers is obtained, the geometric distance between each supplier and the ideal supplier with the best score in each criterion is computed to determine rank among suppliers over multi-criteria.

In fuzzy TOPSIS (Chen et al. 2006), the fitness (or utility) of the candidate suppliers with respect to each criterion is represented by a fuzzy number using linguistic variables and, hence, the ranking of the suppliers is based on the comparison of their corresponding fuzzy utilities. Following the concept of TOPSIS, the fuzzy positive ideal solution (FPIS) and the fuzzy negative ideal solution (FNIS) are defined and the distance of each supplier from FPIS (D+) and FNIS (D-) are calculated. Finally, the closeness coefficient (CC) value of each supplier is calculated and the ranking order of all candidate suppliers is determined to select the best one from a set of candidate suppliers.

4.3   Framework of the Proposed Agile Supply Chain DSS

The research framework for agile supply chain DSS consists of the following four major stages: (1) identification of evaluation criteria and decision hierarchy, (2) identification of evaluation criteria weights, (3) supplier evaluation and selection using the FAHP and fuzzy TOPSIS methods, (4) business impact assessment of the proposed system. We present the first three stages in the current section, and the last stage, business impact assessment of the proposed system, is presented in the next section.

4.3.1 Identification of Evaluation Criteria and Decision Hierarchy

In the first stage, we identify both quantitative and qualitative evaluation criteria to evaluate candidate suppliers to formulate supplier evaluation and selection problem as an MCDM problem. In particular, we identify several sub-criteria within in each

main criterion and form the decision hierarchy based on relationships between main and sub-criteria toward the ultimate goal of selecting the fittest suppliers. To this end, five managers or assistant managers from a leading Korean automotive company were interviewed as industry experts to identify and determine the necessary criteria. During this interview, the critical factors that influence decision-makers with respect to the core requirements of buyers were identified and synthesized. In addition, we have reviewed the literature to collect the criteria adopted in previous supplier evaluation problems. Specifically, we have searched with the keywords such as 'supplier selection', 'supplier evaluation', 'vendor selection', 'partner selection', 'supply chain design', and so on. Next, we have identified more than 20 studies in leading journals such as European Journal of Operation Research, International Journal of Production Economics, Journal of Purchasing. In the end, four main evaluation criteria, namely the general management capability perspective (GP), manufacturing capability perspective (MP), collaboration capability perspective (CP), and agility perspective (AP) were identified and verified by industry experts in terms of their significance and attainability. Each criterion includes two to nine sub-criteria. Table 4.2 describes a set of 25 sub-criteria of the four main criteria. Note that some sub-criteria, such as production facility and capacity, could either be categorized as a type of manufacturing capability or as a type of agility. In this case, experts are asked to determine which main criterion is the best fit for sub-criteria.

Once the supplier evaluation criteria were identified, the decision hierarchy for supplier selection at four levels was structured as shown in Fig. 4.2. The top level of the hierarchy represents the ultimate goal of selecting the best supplier. The second level is grouped under criterion of GP, MP, CP, and AP perspective. At the third level, various sub-criteria that measure the candidate suppliers in detail are presented. Finally, the bottom level of the hierarchy presents the candidate suppliers.

4.3.2 Identification of Evaluation Criteria Weights

After forming the decision hierarchy, we calculate the weights of the evaluation criteria using the FAHP. To this end, linguistic variables were used by the experts for the pairwise comparisons with fuzzy preferences, and fuzzy values were converted into a crisp number by applying *extent analysis* (Chang 1992). The importance of each criterion was compared using results from a questionnaire, while the preference of one measure over another was decided by experts with respect to the main goal. The geometric means of these values were then calculated to obtain an agreed pairwise comparison. Further, a consistency index and consistency ratios in a crisp AHP were applied to keep the result consistent. Each triangular fuzzy number was also converted into a crisp number and the consistency of each matrix was checked. The consistency ratio of each pairwise comparison matrix was calculated as being under 0.1, which indicated that the weights were consistently reliable.

Table 4.2 Descriptions of the main criteria and sub-criteria

Main criteria	Sub-criteria	Description
GP	Management and strategy (MS)	The degree to which a supplier is in line with the firm's vision, strategy, and policy
	Financial status (FS)	The degree to which a supplier is financially stable
	Customer relations (CR)	The degree to which a supplier has strong customer relationships
	Training program (TP)	The degree to which a supplier has well-defined HR training programs
	Reputation (RE)	The degree to which a supplier has a good reputation
	History (HI)	The degree to which a supplier has a long history in the business
	Language (LA)	The degree to which a supplier has the ability to communicate in multiple languages
	License/Certification/ Award (LCA)	The degree to which a supplier has certified qualifications
	Geographical location (GL)	The degree to which a supplier is located nearby
MP	Production facility/capacity (PFC)	The degree to which a supplier has considerable production capacity
	Product diversity (PD)	The degree to which a supplier offers diversified products
	R&D capability (RD)	The degree to which a supplier puts effort into R&D activities
	Safety regulations (SR)	The degree to which a supplier obeys safety regulations
	Environmental regulations (ER)	The degree to which a supplier is environmentally friendly
	Quality control (QC)	The degree to which a supplier conducts quality control actively
	Product price (PP)	The degree to which a supplier offers a cheaper price
CP	After-sales service (AS)	The degree to which a supplier provides good after-sales service
	Delivery reliability (DR)	The degree to which a supplier meets delivery requirements
AP	Delivery speed (DS)	The degree to which a supplier meets delivery speed requirements
	Delivery flexibility (DF)	The degree to which a supplier can respond to changes in delivery requirements
	Make flexibility (MF)	The degree to which a supplier can respond to changes in production requirements
	Source flexibility (SF)	The degree to which a supplier can respond to changes in source requirements
	Agile customer responsiveness (ACR)	The degree to which a supplier can respond to changes in customer requirements
	Collaboration with partners (CPB)	The degree to which a supplier can collaborate across each partner's core business
	IT infrastructure (IT)	The degree to which a supplier adopts a practical IT system
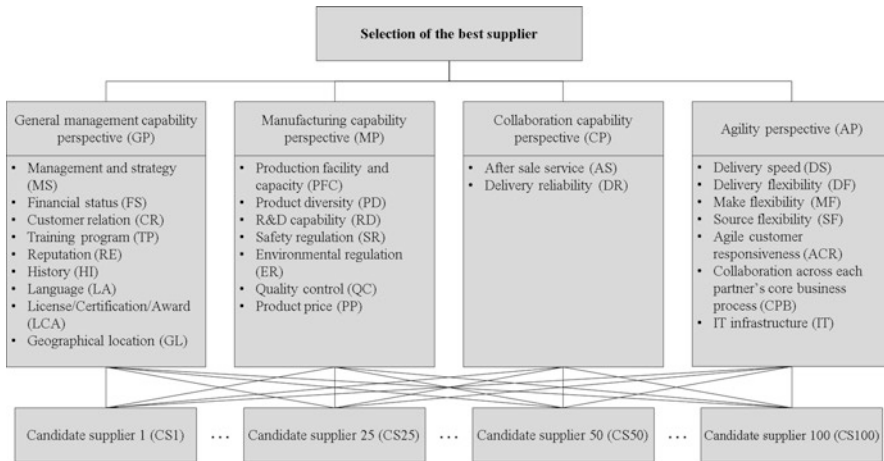
Fig. 4.2 Decision hierarchy of supplier selection

Table 4.3 Priority weights for evaluation criteria

Main criteria	Priority weights	Sub-criteria (priority weights)	
GP	0.12	MS (0.15)	HI (0.07)
		FS (0.16)	LA (0.00)
		CR (0.17)	LCA (0.14)
		TP (0.00)	GL (0.10)
		RE (0.21)	
MP	0.41	PFC (0.22)	ER (0.00)
		PD (0.00)	QC (0.31)
		RD (0.13)	PP (0.34)
		SR (0.01)	
CP	0.10	AS (0.54)	DR (0.46)
AP	0.38	DS (0.00)	ACR (0.27)
		DF (0.09)	CPB (0.10)
		MF (0.19)	IT (0.33)
		SF (0.02)	

Then, the priority weights of the main criteria and sub-criteria were calculated using the FAHP method. We present calculated priority weights of main criteria and sub-criteria in Table 4.3. According to Table 4.3, industry experts consider MP (0.41) the most important main criteria in the process of selecting the best supplier. Also notable is the fact that they estimate the importance of AP to be 38 % in the process of selecting the best qualifying suppler in automotive industry. This is significant considering the fact that experts were not informed of or influenced by any member of research teams to weigh more this new criterion, but were simply advised to determine the priority weights of this new criterion along with three traditional criteria to "select the best supplier" in their business domain.

4.3.3   Supplier Identification, Evaluation, and Selection

We now illustrate an exemplar case of evaluating and selecting an automotive parts supplier to demonstrate the utility of new AP criterion and the proposed DSS. This case study was conducted as part of the *i-Manufacturing* project under the sponsorship of the Korean Ministry of Knowledge Economy initiated to inspire greater innovation into (automotive) manufacturing industries and encourage technology-centered collaborations among small and medium-sized enterprises in Korea. Two different types of data were collected during this project period. The first data set was priority weights of evaluation criteria from five managers working at medium-sized suppliers in automotive industry through formal lengthy interviews followed by rewritten records. The second data was candidate suppliers' information. While a small set of actual suppliers of automotive parts was identified, it was an extremely tedious and time consuming process for industry experts to evaluate each supplier for each criterion. This is mainly because many small and medium-sized suppliers could not provide sufficient digitized information for experts to determine their capability to meet the requirements associated with each criterion. In addition, a small set of available suppliers was not diverse in terms of their capabilities for each criterion enough to rigorously test the proposed concept of the agile supply chain DSS.

As an alternative, we decided to artificially generate a number of suppliers with well distributed evaluation scores over main criteria and sub-criteria. Specifically, we simulated 100 candidate suppliers' information (CS1–CS100) along with virtual experts' judgment of each candidate supplier for the individual criterion using the linguistic variables and their corresponding fuzzy numbers: (0,1,3)-very poor, (1,3,5)-poor, (3,5,7)-fair, (5,7,9)-good, and (7,9,10)-very good. Note also that we maintain positive relationships between judgment values of sub-criteria within the same criterion with approximately 20 % of noise level so that a supplier s_i with the highest value for a criterion c_i is not necessarily receives the highest value for other criteria c_j while the value of c_j is likely to be high. Meanwhile, this study employed the average value method to integrate the fuzzy judgments of each criterion. Then, it becomes straightforward to calculate the distance of each supplier from FPIS (D^+) and FNIS (D^-). Finally, the closeness coefficient (CC) value of each supplier is calculated to determine the ranking order of all candidate suppliers. Rankings of exemplar suppliers are described in Table 4.4.

Table 4.4   Fuzzy TOPSIS results and rankings

	w/ considering AP (agile supply chain)				w/o considering AP (non-agile supply chain)			
	D^+	D^-	CC	Ranking	D^+	D^-	CC	Ranking
CS1	**0.180**	**0.841**	**0.824**	**1**	**0.165**	**0.474**	**0.742**	**3**
CS2	0.246	0.779	0.760	4	0.203	0.438	0.683	13
CS3	0.409	0.625	0.605	25	0.191	0.450	0.702	8
CS4	**0.379**	**0.654**	**0.633**	**21**	**0.083**	**0.554**	**0.870**	**1**

Note that we present two different outcomes of fuzzy TOPSIS in Table 4.4, one that considers all four main criteria and their associated sub-criteria, and the other that consider three main criteria after eliminating AP. We denote the resulting supply chain based on the first outcome of fuzzy TOPSIS as agile supply chain, whereas the resulting supply chain based on the second outcome as non-agile supply chain mainly because it does not consider AP criterion. Therefore, the first outcome of fuzzy TOPSIS in Table 4.4 was based on the absolute values of priority weights of main criteria and sub-criteria in Table 4.3, the second outcome was based on the adjusted values of priority weights by using the relative ratios of priority weights of three main criteria.

According to Table 4.4, the ranking orders of the candidate suppliers were very different depending whether or not AP criterion was indeed considered as an input of fuzzy TOPSIS. For example, based on the value of *CC*, CS1 was chosen as the best supplier when AP was considered, while CS4 was the best when AP was not considered. In addition, while CS1 was a top ranked supplier in either cases (i.e., 1st with AP and 3rd without AP), rankings of many other suppliers in both cases were significantly different (e.g., CS3 was ranked 25th with AP, but 8th without AP). This in turn insinuates that the resulting agile supply chain with the best supplier with AP and non-agile supply chain with the best supplier without AP consideration will be very different in their nature of meeting the business requirements. In the following section, we compare business impacts of two resulting supply chains in a very simple configuration.

4.4 Assessing Business Impact of Agile Supply Chain DSS

4.4.1 Simple Supply Chain Configuration

The business impact of the proposed agile supply chain DSS is quantified by comparing the magnitude of bullwhip effect of the resulting agile supply chain suggested by the proposed DSS that explicitly considers agility criterion with that of the legacy supply chain system that are partially agile or not agile at all. Note that bullwhip effect is the phenomenon in which variance of demand information is amplified when moving upstream in supply chains (Lee et al. 1997). To quantify the magnitude of bullwhip effects of agile and non-agile supply chains, we imagine a very simple supply chain consisting of one buyer and one supplier, and the customer demand follows AR(1) autoregressive model. In this simple supply chain, order lead time is considered to be fixed and the buyer is assumed to employ a simple order-up-to inventory policy with the forecasted demand based on either minimizing mean-squared forecast error technique or moving average forecasting method. We adopt the following set of notations to describe this simple supply chain.

D_t	Demand of period t
q_t	Order quantity at the beginning of period t
S_t	Order-up-to level at the beginning of period t
ϕ	The first-order autocorrelation coefficient
δ	Constant of the autoregressive model
μ_d	Mean of the autoregressive process which is used to describe the demand process
σ_d^2	Variance of demand
L	Order lead time
D_t^L	Lead-time demand
$\hat{\sigma}_t^L$	Standard deviation of lead-time demand forecast error
ε_t	Forecast error for period t, *i.i.d.* from a normal distribution with mean 0 and variance σ^2
z	Constant chosen to meet a desired service level

Note that S_t, the inventory position at the beginning of period t after the order has been placed, is calculated as $S_t = \hat{D}_t^L + z\hat{\sigma}_t^L$ after considering the lead-time demand forecast of period t to support the desired service level in accordance to the business strategy. Once the value of S_t is determined, the order quantity, q_t, at the beginning of period t is determined as follows: $q_t = S_t - S_{t-1} + D_{t-1}$. Since we assume that the customer demand follows AR(1) autoregressive model, it is modeled by $D_t = \delta + \phi D_{t-1} + \varepsilon_t$, and it is trivial to show that $E(D_t) = \mu_d = \dfrac{\delta}{1-\phi}$ and $Var(D_t) = \sigma_d^2 = \dfrac{\sigma^2}{1-\phi^2}$. Now, if the buyer forecast demand by minimizing the expected mean squares of error, lead time demand forecast and its variance can be determined by

$$\hat{D}_t^L = \mu_d\left(L - \frac{\phi\left(1-\phi^L\right)}{1-\phi}\right) + \frac{\phi\left(1-\phi^L\right)}{1-\phi}D_{t-1} \text{ and } \left(\hat{\sigma}_t^L\right)^2 = \frac{\sigma_d^2\left(1+\phi\right)}{1-\phi}\sum_{i=1}^{L}\left(1-\phi^i\right)^2,$$

respectively. Then, order quantity and its variance is obtained by

$$q_t = \frac{1-\phi^{L+1}}{1-\phi}D_{t-1} - \frac{\phi\left(1-\phi^L\right)}{1-\phi}D_{t-2} \text{ and } VAR(q_t) = \frac{\left(1+\phi\right)\left(1-2\phi^{L+1}\right)+2\phi^{2L+2}}{1-\phi}\sigma_d^2,$$

respectively (Luong 2007). Finally, the magnitude of bullwhip effect (B) can be determined by taking the ratio of the variance of order quantity and the variance of demand as shown in Eq. 4.1 (Chen et al. 2000; Luong 2007).

$$B = \frac{Var(q_t)}{\sigma_d^2} = \frac{\left(1+\phi\right)\left(1-2\phi^{L+1}\right)+2\phi^{2L+2}}{1-\phi}. \tag{4.1}$$

On the other hand, if demand is forecasted with moving average forecasting of p observations, the magnitude of bullwhip effect can be calculated as a lower bound form (Chen et al. 2000) as shown in Eq. 4.2:

$$B = \frac{Var(q_t)}{\sigma_d^2} \geq 1 + \left(\frac{2L}{p} + \frac{2L^2 + z^2L}{p^2}\right)\left(1-\phi^p\right). \tag{4.2}$$

4.4.2 Bullwhip Effect with Autocorrelation Coefficient and Desired Service Level

We compare the magnitude of bullwhip effect of agile supply chain suggested by the proposed DSS that explicitly considers agility criterion with that of the legacy (or non-agile) supply chain system. In this section, we are particularly interested in estimating the magnitude of bullwhip effect that changes in the first-order autocorrelation coefficient of the customer demand and desired service level may bring. By definition, with higher values of autocorrelation coefficient, the demand of the current time is more dependent on the demand of the last time, and the demand line is fluctuating more violently due to the increasing variance. Therefore, it is easy to conjecture that the magnitude of bullwhip effect of two supply chains increases as the value of autocorrelation coefficient increases. However, since agile supply chain should be able to adapt much swiftly to unexpected customer demand than non-agile supply chain, we expect the difference in the magnitude of bullwhip effect of two supply chains becomes larger as the value of autocorrelation coefficient increases.

We show the magnitude of bullwhip effect of agile and non-agile supply chains in Fig. 4.3a, where x-axis and y-axis represents the first-order autocorrelation coefficient in demand (ϕ) and the magnitude of bullwhip effect (B) computed by Eq. 4.1, respectively. As expected, the magnitude of bullwhip effect from agile supply chain with the most agile supplier is much lower than that of non-agile supply chain with a supplier chosen without considering the agility although the differences in B values of two supply chains are greatly affected by ϕ. First of all, the maximum magnitude of bullwhip effect of agile supply chain does not exceed three, indicating that the variance of order quantity is less than three times that of demand. However, in non-agile supply chain, the maximum value of B is greater than nine, reflecting much greater fluctuation in variance of order quantity over variance of demand. Also our data sets confirm that the lead time of agile and non-agile supply chain is one and nine days, respectively. Second, the difference of B values of two supply chains is not noticeable until $\phi = 0.3$, but since then, the difference becomes very obvious with higher values of ϕ, implying that keeping supply chain agile becomes even more critical when demand pattern is highly correlated with last demand pattern to minimize unexpected risks estimated through B values. Interestingly, B values of two supply chains start to decrease after ϕ reaches a certain point (i.e., $\phi = 0.7$ and $\phi = 0.9$ for agile and non-agile supply chain, respectively) partly because it is relatively easy for the buyer to forecast highly correlated demand. Overall, these observations imply that the buyer with non-agile supply chain is likely to suffer from much larger bullwhip effect, and find it difficult to stabilize the production process. In our following analyses, we limit our discussion to the specific value of autocorrelation coefficient ($\phi = 0.8$) for both supply chains.

Similarly, we illustrate the relationship between the value of B and desired service level (z) in Fig. 4.3b. Note that service level used in calculating order-up-to-level is strategically determined by the decision maker of the buyer in supply chain and it
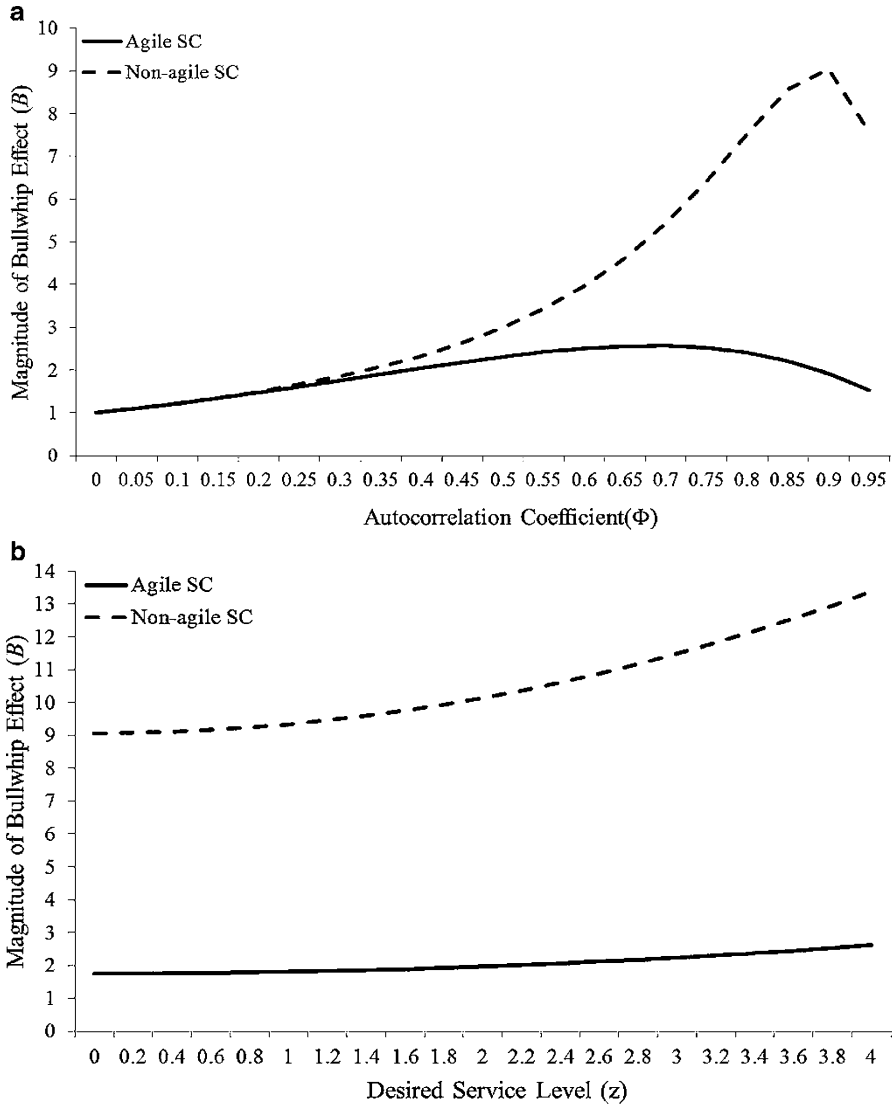
Fig. 4.3 Bullwhip effect trends with autocorrelation and service level. (**a**) Autocorrelation coefficient. (**b**) Desired service level

ultimately determines stockout (or out-of-stock) rate. Since frequent stockouts cause lost sales, dissatisfy shoppers, and diminish store loyalty, it is important to maintain an appropriate service level while avoiding extremely high inventory costs incurred from overstocks. In our simple supply chain, when $z=0$, stockout rate is estimated to be 15.8 %, and when $z \geq 3$, stockout rate is less than 0.1 %. Note that Eq. 4.1 does not include desired service level as an input parameter and hence the value of B does

not change even if the desired service level varies. Therefore, we use Eq. 4.2 to determine the lower bound of B assuming that the normally distributed demand is forecasted with a moving average of $p = 5$ observations. According to Fig. 4.4, the B values of both agile and non-agile supply chains steadily increase as the buyer seeks for higher service levels mainly because it increases both order-up-to-level and order quantity. However, the trend of B values of agile supply chain is much more stable in respond to the service level than that of non-agile supply chain. While the B values of agile supply chain do not change significantly over various service levels, the B values of non-agile supply chain change significantly from nine when service level is zero chain by seven to close to 14 when service level is four. Therefore, when the buyer intends to pursue higher service, she needs to revise her non-agile supply chain into agile supply chain so that she can effectively minimize the negative impacts from bullwhip effects.

4.4.3 The Pareto Fronts with Various Weights of Agility Criterion

Figure 4.4 illustrates the Pareto fronts of two supply chains. The Pareto front of agile supply chain (shown as solid line) represents B values of agile supply chain constructed with a set of 1st ranked suppliers as we vary the weight of AP evaluation criteria between 0 and 1 while maintaining the same relative weight ratios among three other main criteria and sub-criteria within each main criteria as in Table 4.3. That is, the Pareto front of agile supply chain simply displays minimized B values of agile supply chain associated with 1st ranked suppliers that naturally vary as we increase or decrease the weight of AP in the evaluation process of candidate suppliers.

Fig. 4.4 The pareto fronts of agile and non-agile supply chains

Note, however, that *B* values of non-agile supply chain do not change even if we vary the weight of AP because, by definition, non-agile supply chain is constructed with 1st ranked supplier chosen without considering AP criterion. Therefore, the same supplier will be always chosen and, hence, the Pareto front of non-agile supply chain should be repr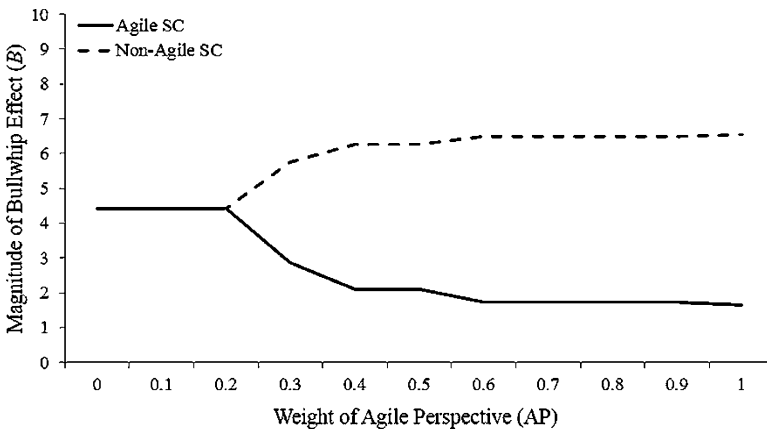esented as a straight line. In order to obtain dynamically changing Pareto front of non-agile supply chain without sacrificing the applicability of the proposed model and the generalizability of managerial insights, we first compute the ratio of lead time of agile to non-agile supply chains for a chosen weight of AP criterion. Since 1st ranked supplier and the corresponding lead time of agile supply chain varies as the weight of AP criterion varies, this ratio dynamically changes as the weight of AP criterion varies. Specifically, the ratio is supposed to increase at a decreasing rate as the weight of AP criterion increases because the lead time of agile supply chain with higher agility decreases. Once we obtain the ratio of lead time over various weights of AP criterion, we multiply it with the original fixed lead time of 1st ranked supplier of non-agile supply chain, and calculate *B* values from the lead time to obtain dynamically adjusted Pareto front of non-agile supply chain. The dotted concave line in Fig. 4.4 represents such Pareto front of non-agile supply chain. Two Pareto fronts intersect at the point where AP weight is set to zero.

As expected, the Pareto front of agile supply chain shows the trend of decreasing *B* values because suppliers with more agility are chosen as 1st ranked suppliers as the weight of AP increases. One of managerial implications obtained from the Pareto front of agile supply chain is that it potentially provides the list of best suppliers as well as corresponding *B* values of the supply chain that the buyer can choose based on her unique and subjective business strategy. For example, when the buyer does not consider agility as critical criterion for configuring her supply chain, the magnitude of bullwhip effects that her best configured supply chain faces will follow the Pareto front of non-agile supply chain. However, if she review and restructure her supply chain into agile supply chain to effectively respond to unexpected demand and market fluctuations, the magnitude of bullwhip effects of the buyer will follow the Pareto front of agile supply chain. Therefore, if we assume that the buyer reconfigures her supply chain from non-agile supply chain to agile supply chain by significantly increasing the weights of AP from 0 to 0.4, the magnitude of bullwhip effect of her agile supply chain drops 4.5–2.1 (i.e., 53 % decrease) following the path of the Pareto front of agile supply chain. However, this magnitude of positive business impact is regarded as an overly conservative estimate considering the fact that if she does not reconfigure her supply chain into agile supply chain by ignoring the importance of agility in fluctuation market situations, her supply chain will suffer from greater bullwhip effects along the Pareto front of non-agile supply chain. In such a case, the *B* value that her non-agile supply chain will face increases from 4.5 to 6 (i.e., 33 % increase). Taken all together, the total business impact of reconfiguring supply chain by increasing the relative weight of agility in supplier selection is the sum of business impacts due to the increased agility (53 %) and the nature of the new agile supply chain (33 %).

Another important managerial implication that decision makers obtain from Fig. 4.4 is that she can determine an appropriate weight of agility in her supply

chain configuration. That is, while it is tempting to make her supply chain extremely agile, the positive business impact due to decreased *B* values increases at a decreasing rate. For example, even if the buyer decides to increase the weight of AP from 0.4 to 1.0, *B* values only marginally change and hence she may not enjoy any considerable business impacts. In contrast, the buyer may enjoy great business impact by increasing the weight of AP from 0.2 to 0.4 easily. This way, decision makers of the buyer can make informed judgment to determine an appropriate level of agility in her supply chain configuration.

4.4.4   The Pareto Fronts with Various Weights of Agility Sub-Criteria

We present in Fig. 4.5 outcomes of sensitivity analyses of a sub-criterion within AP while maintaining the priority weight of AP at 0.38 and the same weight ratios among other sub-criteria within AP as in Table 4.3.

For example, Fig. 4.5a illustrates the Pareto front of agile supply chain as we vary the weight of IT infrastructure (IT) within AP from zero to 1 without changing weight ratios of other sub-criteria. We immediately note that the impact of IT sub-criterion on the *B* value is marginal (but negative) even with its greatest possible weight (1.0) considering the fact that IT (weight 0.33) is considered the most important sub-criterion by experts. The Pareto front of agile supply chain based on adjusted weights of delivery speed (DS) also shows a very similar pattern of IT criterion and hence is not shown.

Figure 4.5b shows that the Pareto fronts of three other sub-criteria such as agile customer responsiveness (ACR), make flexibility (MF), and collaboration across partner's core business process (CPB) all show similar patterns, decreasing trends of *B* values as their weights become greater than certain values (e.g., 0.2 for ACR, 0.1 for MF and CPB). Finally, Fig. 4.5c, d show Pareto fronts of agile supply chains based on varying weights of source flexibility (SF) and delivery flexibility (DF), increasing trends of *B* values as weights of two criteria increase. All these trends in Fig. 4.5a~, d result from the fact that a different supplier is selected as the weight of a chosen sub-criterion varies. One interesting managerial insight gained from Fig. 4.5 is that as priority weights of sub-criteria with minimal weights in Table 4.3 deviate farther from pre-determined weights by industry experts, *B* values greatly increase. For example, as priority weights of DF and SF increase from (0.09 to 0.6) and (0.02 to 0.6), *B* values of agile supply chains significantly increase from 2.1 to 4.5 (114 %), implying that decision makers should not try to significantly vary priority weights of sub-criteria within AP from the suggested weight values by industry experts. Rather, they are strongly suggested to maximize the positive impact of their supply chains by increasing the weight of AP while maintaining similar relative weights of sub-criteria within AP as shown in Table 4.3.
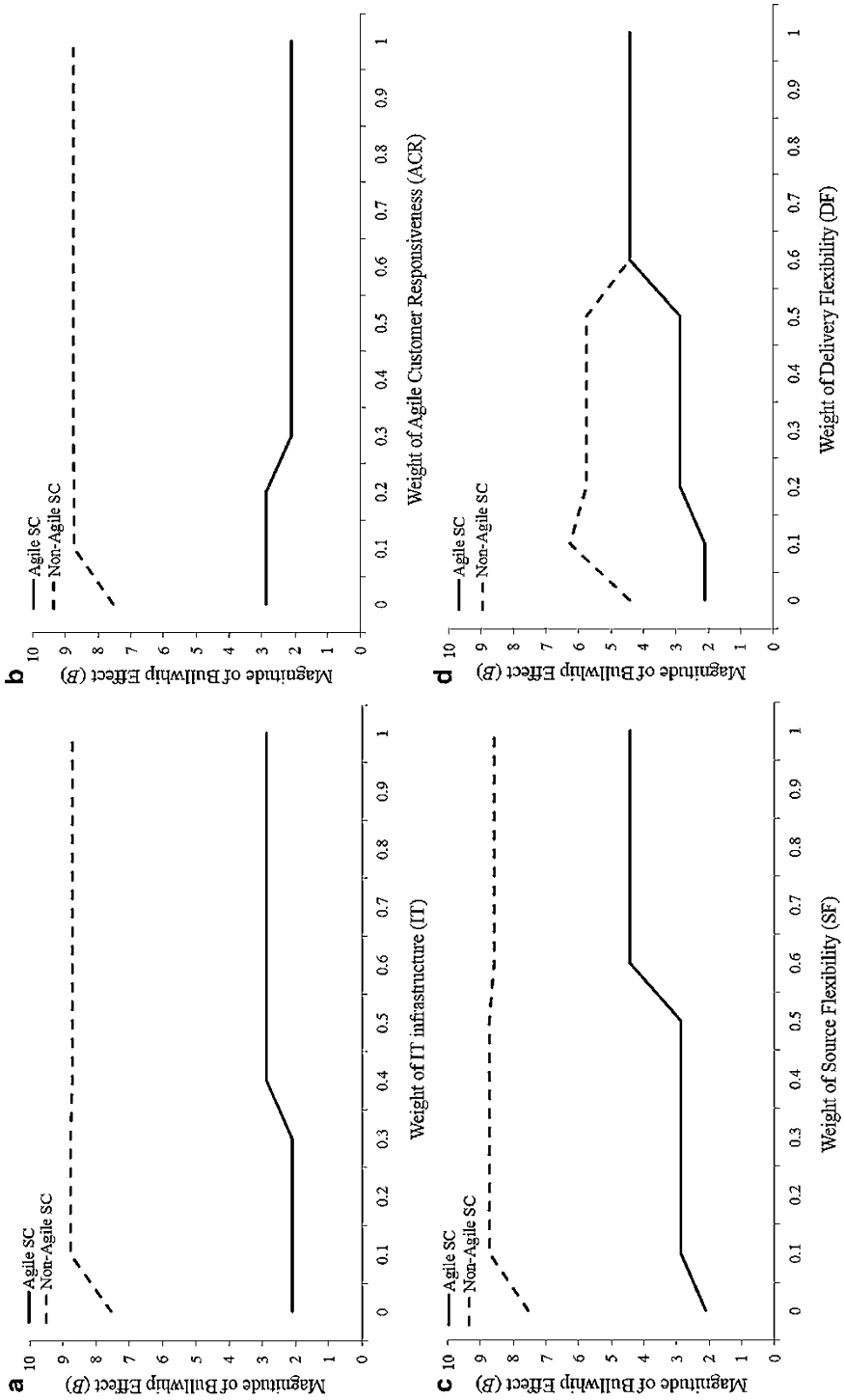
Fig. 4.5 Business impacts of agility sub-criteria (**a**) IT infrastructure. (**b**) Agile customer responsiveness. (**c**) Source flexibility. (**d**) Delivery flexibility

4.5   Conclusion

Obtaining a set of the fittest agile suppliers based on decision makers' unique and subjective multi-criteria is critical to reconstruct a current supply chain into an agile supply chain to significantly improve its responsiveness to unexpected demand fluctuations and maintain market competitiveness. However, selecting the fittest agile suppliers involves sophisticated evaluation processes that require the consideration of several alternatives and sometimes conflicting criteria with a large amount of subjective or ambiguous information. This study presents a fuzzy MCDM methodology with fuzzy AHP and fuzzy TOPSIS to determine relative importance of multi-criteria and assess potential suppliers using triangular fuzzy numbers to translate the subjective judgments of evaluators based on ambiguous linguistic variables into quantifiable numeric estimates. Most importantly, we find that agility criterion in supplier selection is viewed as the second most important criterion by experts in automotive industry in Korea and two resulting supply chains, with and without agility criterion, show dramatically different performance measured by the magnitude of bullwhip effect. Overall, building and maintaining agile supply chain becomes more important when the current demand is highly correlated with past demand pattern and/or high service level is strategically pursued. In particular, Pareto fronts of two supply chains not only graphically show the total business impact obtained from reconfiguring the current supply chain by increasing the relative weight of agility in supplier selection but also help decision makers determine an appropriate level of agility in her supply chain configuration by considering a decreasing growth rate of positive business impact as a higher level of agility is sought.

Also, the agile supply chain building outcomes discussed in this study provide direction to managers in the evaluation criteria that should be developed to assess the success or failure of supply chain risk management. As agile supply chain helps the firms accomplish rapid adaptation of uncertainties in the operating environment, the proposed decision support system could be an important initial step in developing models of supply chain risk management strategies. Practically, managers could understand the advantages and estimate business impact of their decision makings to assess and manage the myriad of supply chain uncertainties.

Biography

Jaehun Lee   is currently a Ph.D. candidate in the Department of Industrial and Management Engineering at the Pohang University of Science and Technology (POSTECH). He received his B.S. degree from POSTECH in 2010. His research interests include supply chain management, business systems analysis and innovation, and marketing.

Hyunbo Cho   is a professor in the Department of Industrial and Management Engineering at the Pohang University of Science and Technology (POSTECH). He

received his B.S. and M.S. in Industrial Engineering from Seoul National University in 1986 and 1988, respectively, and his Ph.D. in Industrial Engineering with specialization in Manufacturing Systems Engineering from Texas A&M University in 1993. He is a recipient of the SME's 1997 Outstanding Young Manufacturing Engineer Award. His areas of expertise include business systems analysis and innovation, and supply chain management. He is an active member of IIE and SME.

Yong Seog Kim   is an associate professor in Management Information Systems department at the Utah State University. He received his M.S. degree in Computer Science and Ph.D. in Business Administration from the University of Iowa. Dr. Kim's primary research interest is in decision support systems utilizing various data mining (KDD) algorithms such as variable selection, clustering, classification, and ensemble methods. His papers have appeared in Management Science, Decision Support Systems, Intelligent Data Analysis, Expert Systems with Application, and Journal of Computer Information Systems, and conference proceedings of KDD, AMCIS, DSI, HICSS, and many others. Dr. Kim currently serves on the editorial board of the Journal of Computer Information Systems, Journal of Information Technology Cases and Applications, and Journal of Emerging Trends in Computing and Information Sciences.

References

Araz, C., Ozfirat, P. M., & Ozkarahan, I. (2007). An integrated multicriteria decision-making methodology for outsourcing management. *Computers & Operations Research, 34*(12), 3738–3756.

Bottani, E., & Rizzi, A. (2008). An adapted multi-criteria approach to suppliers and products selection—An application oriented to lead-time reduction. *International Journal of Production Economics, 111*(2), 763–781.

Bray, R. L., & Mendelson, H. (2012). Information transmission and the bullwhip effect: An empirical investigation. *Management Science, 58*(5), 860–875.

Chan, F. T. S., Kumar, N., Tiwari, M. K., Lau, H. C. W., & Choy, K. L. (2008). Global supplier selection: a fuzzy-AHP approach. *International Journal of Production Research, 46*(14), 3825–3857.

Chang, D.-Y. (1992). Extent analysis and synthetic decision. *Optimization Techniques and Applications, 1*(1), 352–355.

Chen, F., Drezner, Z., Ryan, K. J., Ryan, K. J., & Simchi-levi, D. (2000). Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, lead times, and information. *Management Science, 46*(3), 436–443.

Chen, C.-T., Lin, C.-T., & Huang, S.-F. (2006). A fuzzy approach for supplier evaluation and selection in supply chain management. *International Journal of Production Economics, 102*(2), 289–301.

Christopher, M. (2000). The agile supply chain: Competing in volatile markets. *Industrial Marketing Management, 29*(1), 37–44.

Christopher, M., & Towill, D. (2001). An integrated model for the design of agile supply chains. *International Journal of Physical Distribution and Logistics Management, 31*(4), 235–246.

Ho, W., Xu, X., & Dey, P. K. (2010). Multi-criteria decision making approaches for supplier evaluation and selection: A literature review. *European Journal of Operational Research, 202*(1), 16–24.

Hwang, C. L., & Yoon, K. (1981). *Multiple attribute decision making methods and applications: A state-of-the-art survey*. New York: Springer.

Kim, Y., Street, W. N., Russell, G. J., & Menczer, F. (2005). Customer targeting: A neural network approach guided by genetic algorithms. *Management Science, 51*(2), 264–276.

Klein, R., & Rai, A. (2009). Interfirm strategic information flows in logistics supply chain relationships. *MIS Quarterly, 33*(4), 735–762.

Krishnamurthy, R., & Yauch, C. (2007). Leagile manufacturing: A proposed corporate infrastructure. *International Journal of Operations and Production Management, 27*(6), 588–604.

Kumar, M., Vrat, P., & Shankar, R. (2004). A fuzzy goal programming approach for vendor selection problem in a supply chain. *Computers & Industrial Engineering, 46*(1), 69–85.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*(4), 546–558.

Lee, H. L., So, K. C., & Tang, C. S. (2000). The value of information sharing in a two-level supply chain. *Management Science, 46*(5), 626–643.

Lin, C.-T., Chiu, H., & Chu, P.-Y. (2006). Agility index in the supply chain. *International Journal of Production Economics, 100*(2), 285–299.

Luong, H. (2007). Measure of bullwhip effect in supply chains with autoregressive demand process. *European Journal of Operational Research, 180*(3), 1086–1097.

Önüt, S., Kara, S. S., & Işik, E. (2009). Long term supplier selection using a combined fuzzy MCDM approach: A case study for a telecommunication company. *Expert Systems with Applications, 36*(2), 3887–3895.

Power, D. J., Sohal, A. S., & Rahman, S. U. (2001). Critical success factors in agile supply chain management-An empirical study. *International Journal of Physical Distribution & Logistics Management, 31*(4), 247–265.

Raharjo, H., Brombacher, A. C., & Xie, M. (2008). Dealing with subjectivity in early product design phase: A systematic approach to exploit Quality Function Deployment potentials. *Computers & Industrial Engineering, 55*(1), 253–278.

Saaty, T. L. (1980). *The analytic hierarchy process*. New York: McGraw-Hill.

Wallenius, J., Dyer, J. S., Fishburn, P. C., Steuer, R. E., Zionts, S., & Deb, K. (2008). Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Management Science, 54*(7), 1336–1349.

Yao, Y., & Zhu, K. X. (2012). Research note—Do electronic linkages reduce the bullwhip effect? An empirical analysis of the US manufacturing supply chains. *Information Systems Research, 23*(3), 1042–1055.

Yusuf, Y. (2004). Agile supply chain capabilities: Determinants of competitive objectives. *European Journal of Operational Research, 159*(2), 379–392.

Chapter 5
Hawkes Point Processes for Social Media Analytics

Amir Hassan Zadeh and Ramesh Sharda

Abstract Online social networks (OSNs) produce a huge volume of content and clickstream data over time as a result of continuous social interactions between users. Because these social interactions are not fully observable, the mining of such social streams is more complex than traditional data streams. Stochastic point processes, as a promising approach, have recently received significant research attention in social network analysis, in attempts to discover latent network structure of online social networks and particularly understand human interactions and behavior within the social networks. The objective of this paper is to provide a tutorial to the point process framework and its implementation in social media analytics. It begins by providing a quick overview of the history of Hawkes point processes as the most widely used classes of point process models. We identify various capabilities and attributes of the Hawkes point processes and build a bridge between the theory and practice of point processes in social network analytics. Then the paper includes a brief description of some current research projects that demonstrate the potential of the proposed framework. We also conclude with a discussion of some research opportunities in online social network and clickstream point process data.

Keywords Point process • Hawkes process • Self-exciting • Mutual-exciting • Online social networks • Twitter • Online content • Stream mining

5.1 Introduction

During the past few years, millions of people and organizations have used online social networks applications (Facebook, Twitter, YouTube, Google+, etc.) as a part of their daily online activities (Guy et al. 2010). As a result of continuous social interactions between participants over these websites, these platforms have generated, and will continue to generate, enormous amount of data over time. Understanding

A. Hassan Zadeh • R. Sharda (✉)
Department of Management Science and Information Systems,
Oklahoma State University, Stillwater, OK, USA
e-mail: ramesh.sharda@okstate.edu; http://spears.okstate.edu/profiles/?id=71

rules and structures governing human interactions and collective behavior is a challenging task in the field of social network analysis. This paper is concerned with the latent networks that cannot fully be observed in online social networks, but have to be analyzed for different purposes. For example, videos on YouTube are watched thousands of times; tweets on Twitter are retweeted, replied and marked as favorite many times by followers; Wikipedia pages are edited quite frequently by contributors; online advertisements or brand posts on Facebook are clicked on by users resulting in popping a web page up from which a purchase may be completed. The common thing among these examples is that activities on one piece of information are likely to cause follow-up activities on itself and other related content. How can we elucidate such interactions and similarities, unravel them from massive unstructured data that generate throughout the days and leverage them for businesses purposes? Numerous approaches have been proposed to address this problem in different ways. Point processes are one of these approaches that have recently received substantial attention in social and behavioral sciences due to their ability to resemble the dynamics of social interactions and discover hidden patterns and implicit network structure within the underlying social structures. On the other hand, the availability of time-stamped and geo-located data collected by various data technologies over these platforms have made it possible to treat this data like point process data, work with point process models and study the spatial-temporal properties of the dynamic social networks.

In this paper, we consider a class of point process models capable of resembling the temporal pattern of social interactions and specifying the temporal dependencies of the interaction events with a branching structure within online social networks. Hawke's self-exciting and mutual-exciting point process (Hawkes 1971) and Ogata's ETAS (Epidemic Type After shock Sequences) model (Ogata 1988), an extension of Hawkes process, are very flexible frameworks which best suited for highly dynamic networks. The Hawkes process model was first designed to model the branching structure of earthquakes. It implemented the idea that an earthquake can trigger aftershocks. Later, the Ogata's point process model was developed to implement the idea that aftershocks sequences have an epidemic behavior, i.e. a large earthquake induces more aftershocks than small earthquakes do. An earthquake with a large magnitude triggering aftershocks is analogous to the activity of an influential user on a content causing follow-up activities by other users on his/her network of followers. Therefore, adapting a point process framework similar to the Hawkes' self-exciting and mutual-exciting point process and ETAS model can be beneficial to analysis that we wish to perform within online social networks.

To the best of our knowledge, there are relatively few studies in literature that have used the point process framework to study dynamic online social networks (Crane and Sornette 2008; Lawrence and Michael 2010; Steeg and Galstyan 2012; Hassan Zadeh and Sharda 2014a, b; Yang and Zhao 2013). Building upon these studies, this paper provides a tutorial for a point process framework and its implementation to social media analytics. We first briefly provide a quick overview of the point process, then present self and mutually exciting or Hawkes processes in detail.

This section basically reviews the statistical theory underlying the Hawkes point process approach. In Sect. 5.4, we briefly review some of the applications of the Hawkes process across various areas of research including applications not only in seismology and finance, but also in healthcare and bioinformatics, social networks, social media, sociology, criminology and terrorism. We then explain some current research projects that demonstrate the usefullness of the Hawkes point process and how the point processes framework is mapped out, to some extent, in online social networks. The final section presents a general conclusion of this paper.

5.2 Point Processes

A point process is a type of random process or a stochastic mechanism that can generate times and spaces of events of a point pattern which we may observe. For instance, the occurrence of an earthquake, fire, neuronal spikes, crime or new thread, content and conversation on social media might be regarded as a point process in both time and geographical space (or in even more general dimensions) if every single point is recorded and mapped according to its position either in time or space or both. Point process models have long been used for describing such real-world phenomena occurring at random locations and/or times (Schoenberg et al. 2010). Temporal Point process models are closely rooted in survival analysis which deals with the durations of time between consecutive events. However, point process models focus on times of events that may appear on the timeline (Rathbun et al. 2006).

A point process can be viewed in terms of a list of times $t_1, t_2, \ldots, t_n$ at which corresponding events *1, 2,…, n* occur (Daley 2006; Daley and Vere-Jones 2003a, b). Intuitively, a point process is characterized by its conditional intensity $\lambda(t)$ which represents the mean spontaneous rate at which events are expected to occur given the history of the process up to time *t* (Ogata 1988). In particular, a version of the conditional intensity may be given by the process

$$\lambda(t) = \lim_{\Delta t \to \infty} \frac{E\left(N[t, t + \Delta t] \middle| H_t\right)}{\Delta t}$$

where H_t denotes the history of events prior to time *t*, and the expectation represents the number of events $N[t, t + \Delta t]$ occurring between time *t* and $t + \Delta t$. The Poisson process is the prototype of a point process that yields random events in space or time where two consecutive events are independent. In other words, a point process is classified as a Poisson process if events occurring at two different times are statistically independent of one another, meaning that an event at time t_1 neither increases nor decreases the probability of an event occurring at any subsequent time. A Poisson process is governed by a single parameter or Poisson intensity. Although Poisson processes have many nice properties which make them particularly well suited for special purposes, they cannot capture interaction effects between events. In the next section, we turn our attention to a more general point process rather than the

stationary Poisson process known as Hawkes process. Very useful sources of theoretical discussions and empirical applications of various types of point processes can be found in the textbooks (Daley 2006; Daley and Vere-Jones 2003a, b).

5.3 Hawkes Point Processes

The origin of Hawkes point process goes back to the seventies when Hawkes was looking for a mathematical model that describes earthquake occurrences. Hawkes (1971) introduced a new class of point process to model earthquake occurrences. What was new in his perspective in contrast to other concurrent approaches was a concrete and mathematically traceable point process model with inclusions of branching, self-exciting and self-similarity behaviors (Liniger 2009). The Hawkes process originally states that when an event occurs, it will increase the chance of occurrence of some future events. Over the past few years, Hawkes process models have received significant attentions from researchers, especially in seismology research in terms of theoretical and empirical implications. Though, there exists several equivalent forms in literature in which Hawkes point process can be defined, standard Hawkes process can be defined as a temporal point process with long memory, branching effect and self-exciting properties. Hawkes process is originally characterized by its associated conditional intensity process which allows us to describe the underlying dynamics of the process in a convenient way. The intensity $\lambda(t)$ at a given time t corresponds to the chance of occurrence of an event between time t and $t + \Delta t$ given the history of the process up to time t. Let $X = \{(t_i, m_i)\}$ to be a marked point process on the timeline, where $t_i \subseteq R$ is an event of the point process and $m_i \subseteq M$ denotes the corresponding mark. The conditional intensity function of the standard Hawkes process is assumed to be of the form

$$\lambda(t) = \mu(t) + \sum_{t_i < t} \alpha(\kappa_i) \beta(t - t_i, m_i) \tag{5.1}$$

Where $\mu(t)$ denotes an immigrant intensity function. Function $\alpha(\kappa)$ represents total offspring intensity and function $\beta(t, \kappa)$ is a density function on $[0, \infty)$. This function is also called normalized offspring intensity which is allowed to depend on the mark m. They are conditional on past events and marks given by history of the process i.e. $H_t : \left\{(t_i, m_i)\right\}_{t_i < t}$.

Now, we turn our attention to multivariate marked Hawkes process. In this article, we put the definition of Daley and Vere-Jones (2003) into the prospective as the generalized closed form of the Hawkes point process. Let $N(t) = [N_1(t), \ldots, N_d(t)]$ be a multivariate point process that is the superposition of several univariate point processes of different types $\{N_1(t), \ldots, N_d(t)\}$. $N_j(t) : j = 1, \ldots, d$ denotes the number of points or events type j in the interval $[0, t)$. By definition, a multivariate marked Hawkes process is a class of d-dimensional point process which has d intensity processes given by:

$$\lambda_j\left(t\,|\,H_t\right)=\mu_j+\sum_{k=1}^{d}\alpha_{kj}\int_{(-\infty,t)\times R}h_{kj}\left(t-s\right)g_j\left(m\right)dN_j\left(s\right)\qquad j=1,2,\ldots,d\qquad(5.2)$$

where the rate of event type (mark) j, $\lambda_j(t)$, is determined by the accumulative self-and mutual-excitement effects of the past occurrence of events of all types. Any one event triggers an increase in the rate of the entire process including its associated intensity process and the other d-1 marked intensity processes. In other words, every event in one of the components increases the intensity of this and all other components. The functions h_{kj} are called transfer functions (also called response, reaction or decay or triggering function in the literature), which are density functions describing the waiting time (lag) distribution between excited and exciting events. These describe how fast the self- and mutual-excitement effects decay in time. The amount of excitement depends on the magnitude of the mark (type) of the triggering event. The fact that a Hawkes process has an underlying clustering structure appears at parameter α_{kj} which indicates the amount of excitation an event type k contributes to the time path of component j. These branching coefficients reflect the overall behavior of the point process that Hawkes found in his paper (Hawkes and Oakes 1974) to be in the interval [0,1) as the necessary condition for existence.

The most commonly used form of the response function is an exponential decay distribution as follows:

$$h_{kj}\left(t-s\right)=\beta_{kj}e^{-\beta_{kj}(t-s)}\qquad\left(\beta_{kj}>0\right)\qquad(5.3)$$

Based on the Hawkes process, intensity function specified in Eq. 5.2 and the density function specified in Eq. 5.3, the conditional intensity for the type-j point process can be written as:

$$\lambda_j\left(t\,|\,H_t\right)=\mu_j+\sum_{k=1}^{d}\rho_{kj}\sum_{\substack{n=1\\\{t_i<t\}}}^{N_k(t)}\beta_{kj}e^{-\beta_{kj}\left(t-t_n^k\right)}\qquad j=1,2,\ldots,d\qquad(5.4)$$

It indicates that if an event has occurred at time s, the intensity is increased at time t by amount of $h(t-s)$. The functions $g_j(m)$ are so-called boost functions which are a distinct feature of marked Hawkes point process describing the strength of the event. In other words, if an event type j with mark x occurs at time t, the effect of this event on the time path of the component j is proportional to $g_j(m)$. While the transfer and boost functions deal with the relative effects of an event, the branching coefficients imply the absolute influence of the event to the timeline (Steeg and Galstyan 2012).

As pointed out earlier, one of the nice properties of Hawkes point process is the ability to handle a branching structure which facilitates incorporating self-excitement, self-excitation and self-similarity behaviors without even taking into consideration the time and the location of the event. This is a different way of relating events to each other in the way as ancestors and offspring are linked together and allows us to bring the theory of branching processes to the context of point process.

Hawkes point processes are commonly fitted with both parametric and non-parametric estimation techniques. Based on the Hawkes process, intensity function written in Eq. 5.1, the likelihood function for any of the individual point process embedded in the entire process can be written as (Daley and Vere-Jones 2003)

$$L = \prod_{k=1}^{d} \prod_{j=1}^{d} \left[\left[\prod_{v=1}^{N_k(t)} \lambda_j \left(t_n^k \mid H_{t_v^k} \right) \right] e^{-\int_0^T \lambda_j (t \mid H_t \, dt)} \right] \qquad (5.5)$$

Numerical maximization algorithms such as the quasi-Newton method, the conjugate gradient method, the simplex algorithm of Nelder and Mead and the simulated annealing procedure are often implemented to compute maximum log-likelihood estimation of Hawkes process models (Daley and Vere-Jones 2003) as there are no analytically close-form solutions available. Veen and Schoenberg (2008) observed that the log-likelihood functions of branching processes of Hawkes type are complex and extremely flat and numerically unstable due to the multidimensionality, incomplete data and hidden network of branching structure of the Hawkes process. They implemented the idea that Hawkes point process data can be viewed as an incomplete data problem in which the unobservable or latent variables ascertain whether an event belongs to a background event or whether it is a foreground event and was triggered by a preceding occurrence. Therefore, they investigated the expectation-maximization (EM) algorithm as an alternative parameter estimation to estimate Hawkes process parameters and found that it is very efficient compared to traditional methods.

The Bayesian nonparametric inference can also be built as an alternative parameter estimation approach for Hawkes process. Rasmussen (2013) implemented an MCMC (Markov Chain Monte Carlo) algorithm i.e. Metropolis-within-Gibbs algorithm, to perform posterior approximations. Usually, a nonparametric approach leads to a more accurate and robust estimation of parameters.

To assess the goodness-of-fit of the fitted conditional intensity, Q-Q plots of the residual process and the durations (the time intervals between the events of residual process) are drawn. The so-called compensator process is used to perform Kolmogorov-Smirnov (K-S) test to assess the reliability of each model as to the extent to which the model fits the data. This criterion provides useful information of the absolute goodness-of-fit of candidate models. Furthermore, the relative ability of each model to describe the data is measured by computing the Akaike information criteria (AIC) (Akaike 1992). The Akaike statistic provides germane numerical comparisons of the global fit of competing models.

When it comes to the simulation of the point process, the thinning method of Ogata (2006) is often used for the simulation of a point process with the estimated intensity function. This method calculates an upper process for the intensity function which is used to simulate a frequency rate on each mark for the time of the next possible event. Then, the ratio of this rate to the upper bound is compared with a uniform distribution to decide whether the simulated occurrence time is accepted or not (Møller and Rasmussen 2005; Harte 2010). This method is also used for the

purpose of prediction. The probability distribution of the time to the next possible event is obtained empirically by simulation outcomes. Based on the in-sample and out-sample performance measures, such as mean absolute error (MAE) (Hyndman and Koehler 2006), the predictive performance of the model can be assessed.

5.4 Hawkes Process Modeling Applications

As mentioned earlier, Hawkes process models have long been used in seismology to recognize similar clustering patterns in earthquake occurrences and to predict subsequent earthquakes, or aftershocks (Adamopoulos 1976; Ogata and Vere-Jones 1984; Ogata 1988; Ogata 1999; Veen and Schoenberg 2008; Wang et al. 2012). In the past few years, point process models have attracted the attentions of researchers from various areas ranging from finance, healthcare and bioinformatics, social networks, to even sociology, criminology and terrorism.

5.4.1 Finance

Starting with papers (Bowsher 2003; Engle and Lunde 2003; Bowsher 2007; Carlsson et al. 2007), Hawkes point process showed the potential to be applicable to a wide variety of problems in economics, finance and insurance. Bowsher (2003, 2007), Engle and Lunde (2003) and later, Bauwens and Hautsch (2009) showed that Hawkes process is able to capture some of the typical characteristics of financial time series. They used Hawkes process models to study the high-frequency price dynamics of financial assets. They proposed mutually exciting or bivariate Hawkes processes as models for the arrival times of trades and succeeding quotes in stock markets and observed that changes in price of a given asset may lead to subsequent quote revisions. Their model helps sellers and buyers determine their pricing strategies by taking into consideration the past prices and trades to decide what price and quote to post. In another study, Hawkes processes have also been proposed as models for the arrival process of buy and sell orders (Carlsson et al. 2007). Bacry et al. (2012a, b) developed multivariate Hawkes process models associated with positive and negative jumps of the asset prices. Their model captures upward and downward changes of prices of assets. Zheng et al. (2014) also extended previous Hawkes process models and proposed a multivariate Hawkes process to describe the dynamics of the bid and ask price of a financial asset.

Another area of finance where Hawkes processes have received significant attentions is risk management and portfolio credit risk analytics. Giesecke and Tomecek (2005), Giesecke et al. (2011) and Errais et al. (2010) revealed that Hawkes processes can also model the credit risk process. They observed that Hawkes processes with exponential transfer function (markov-type Hawkes process) is consistent with the theory of affine jump-diffusion processes in portfolio credit risk and can analyze

price processes for certain credit derivatives. Later, Dassios and Zhao (2012) presented a new point process as a generalization of the cox process with short noise intensity and Hawkes process with exponential decay which combines both self-excited (endogenous) and externally excited (exogenous) factors of the underlying system. They used it to model the credit risk process with the arrival of claims and assumed that bankruptcy is caused by primarily a number of bad events such as credit rating downgrades by rating agencies (endogenous factors) and also other bad news on the company front such as bad corporate financial reports (exogenous factors). Their model is capable of capturing additional aspects of the risk, particularly during the economic downturn which involves plethora of bad economic events. In the same line of research, Chavez-Demoulin and McGill (2012) used Hawkes process models featured by a Pareto distribution for the marks to estimate intraday value-at-risk as one of the important metrics used by market participants engaged in high-frequency trading. In another study, Herrera (2013) applied a marked self-exciting point process model to arrival times of extreme events to estimate value-at-risk in oil markets. These are some among many applications of Hawkes process in finance and other related areas which demonstrate that Hawkes processes have some of the typical characteristics of financial time series.

5.4.2   Healthcare and Bioinformatics

Point process models have also been successfully applied in the analysis of a variety of problems in bioinformatics and healthcare domain. In neurosciences, the spikes are the major components that elicit from real-time information processing in the brain. Brillinger (1975, 1988) and Brillinger et al. (1976) was the first one who proposed the Hawkes process in the field of neurophysiology as a model for neural activity in networks of neurons for understanding the mechanisms of what causes a neuron to spike. They used Hawkes point processes as a tool for identifying the relation between connectivity and spike train correlations in small neural networks. Dahlhaus et al. (1997) used Hawkes processes as a tool for identifying direct and indirect synaptic connections in relatively large neural networks. However, in the following years (till 2010), the literature disregarded the linear Hawkes models perhaps due to nonlinear aspects in spike trains, focusing instead on non-linear point process-type models such as the generalized linear models (GLMs), and related multiplicative models (Cardanobile and Rotter 2010) while not offering the same mathematical exposure and simplicity as Hawkes process does. Krumin (2010), Pernice et al. (2011) and Reynaud-Bouret et al. (2013) used Hawkes process models as a tool for spike train analysis for relating neural spiking activity to spiking history, neural ensemble and exogenous effects. They analyzed effects of different connectivity patterns on correlation structure of neuronal activities and observed that the Hawkes process framework is capable of capturing the dynamics of the spike trains in a linear manner as Hawkes process counterparts.

Recently, Hawkes process has been has been used as an analytical model in human physical activity and health. Paraschiv-Ionescu (2013) proposed Hawkes

process as a model for understanding human physical activity patterns in health and disease, particularly physical behavior in chronic pain. The central question in their research was how chronic pain affects individuals' daily life and movement. They studied the interactions between chronic pain and regular physical activity and observed that Hawkes process is able to capture the temporal dynamics of human activity patterns between periods and events. They concluded that Hawkes process can improve the clinical understanding of chronic pain behaviors by quantifying the complex dynamics of various human activities.

In the area of disease epidemiology, Meyer (2009) and Kim (2011) used Hawkes' self-exciting point process and Ogata's ETAS (Epidemic-Type After shock Sequences) models to study the spread of infectious disease like flu virus during an epidemic or pandemic. They demonstrated that Hawkes process type models can incorporate spatial and temporal dependencies of outbreaks by specifying a branching structure among the outbreaks in order to predict future occurrences of infectious disease and epidemics.

5.4.3 Sociology, Criminology and Terrorism

Hawkes process has been used in many other areas even in sociology, criminology and terrorism. Very recently, several works addressed the potential of the Hawkes-type models to understand and predict future patterns of violent events and security threats. The fact that some crimes, such as burglary and gang violence tend to happen close to each other in both time and space and spread through local environments contagiously, Mohler et al. (2011), Alexey et al. (2011), Hegemann et al. (2013) and Mohler (2013) took advantage of multidimensionality of Hawkes process across time and space as it was implemented in seismology research, studied the behaviors and rivalries of street gangs and observed that Hawkes process and territorial street gangs exhibit similar behavioral characteristics. They used this model to determine the future urban crime hotspots. Porter and White (2012) and Mohler (2013) used Hawkes-type process models to detect terrorist activities and determine the probability of a terrorist attack occurring in a day, location and the severity of the attack. In similar works, the temporal patterns of violent civilization deaths from the Iraq and Afghan conflicts were explored using self-exciting point processes (Erik et al. 2010; Lewis et al. 2012; Zammit-Mangion et al. 2012).

5.4.4 Social Network Analysis

Recently, there has been a growing interest to use Hawkes point process models for social network analysis. Crane and Sornette (2008) and Lawrence and Michael (2010) developed a family of self-exciting point processes to explore the dynamics of viewing behavior on YouTube. They demonstrated that a Hawkes process with a power law response function exhibits similar characteristics of the viral process of

a video on YouTube. These characteristics were classified by a combination of motivational factors (endogenous/exogenous) of user interactions and the ability of viewers to influence others to respond across the network (critical/subcritical). In another study, Lawrence and Michael (2010) used mutually exciting Hawkes process models to understand rules governing collective behaviors and interactions between contributors over Wikipedia. Blundell (2012), Halpin and Boeck (2013) and Masuda (2013) used Hawkes process models to model dyadic and reciprocal interaction within e-mail and conversation networks. Golosovsky (2012) studied the growth patterns of the citation networks and observed that the citation process cannot be a memoryless Markov chain; instead it is consistent with self-exciting point process, since there is an extensive correlation and temporal dependency between the present, recent and past citation rates of a paper. Very recently, Xu et al. (2014) used mutual-exciting Hawkes process to study the dynamic interactions and effect of various types of online advertisements clicks (display, search, purchase) on purchase conversion. Also, Hassan Zadeh and Sharda (2014a, b) and Yang and Zha (2013) built different Hawkes' self- and mutual exciting point process models to investigate the effect of viral diffusion processes on popularity of contents on online social networks.

In summary, Hawkes process has been successfully used in many areas ranging from seismology, finance, medicine, social networks to even criminology and terrorism. Hawkes process models have still this potential to apply to wide variety of other problems to study events or behaviors of interest. Next section outlines two of our previous research projects to demonstrate the capability of Hawkes point process and how point processes models are actually formulated, mapped out and operationalized to some extent in Twitter.

5.5   Hawkes Process Applications in Social Media

Over the past few years, big brands have started taking social media seriously, and social media marketing has been an inevitable part of their marketing plan. As more and more major brands have established their communities within online social networks (OSNs), understanding the behavior of the fans on these platforms became important to the marketers and online content providers in order to enable better organization of online activities, and effective execution of successful marketing campaigns. The central question in our previous works has been to determine the popularity growth patterns of a brand's tweet by analyzing the time-series path of its subsequent activities (i.e. retweets, replies and marks as favorite). Understanding this type of information spreading in social media platforms would potentially allow marketers to predict which trends or ideas will become popular, how fast they will become popular, how much impression a tweet will receive, how long it will be popular, and how often they should tweet. Drawing inspiration from Hawkes process models, we built a self-exciting process model, Ogata's ETAS model and Hawkes mutual-exciting model respectively in order to implement our ideas.

An activity on tweet causing follow-up activities by other users on their network of followers is analogous to an earthquake triggering aftershocks. Second, an earthquake with a large magnitude triggering more aftershocks is analogous to the activity of an influential user on a tweet, inducing more follow-up activities by other users on his/her network of followers; Third, excitation and interaction effects among different types of users' activities (retweets, replies, favorites) is something that needs to be more thoroughly investigated.

The data we crawled from Twitter contained a corpus of an individual brand post tweet, its subsequent activities (retweets, replies, and marks as favorite), along with their timestamps, user IDs and number of followers of the user who contributes to the tweet stream. We took into consideration the timestamp of events, the number of followers and the index or mark attached to it specifying event type "retweet", "reply", and "mark as favorite".

To apply a self-exciting Hawkes process model, we aggregated all users' activities into one single stream of information irrespective to the types of events. As mentioned earlier, self-exciting point process is a simple case of multivariate marked Hawkes point process which is actually a univariate point process, and therefore there is only one intensity process. It indeed ignores the exciting effects among different types of users' activities on a brand's tweet.

To apply an ETAS model, we assumed that the content popularity can be a joint probability function of time and the number of followers. We focused more on incorporating the number of followers as an influential metric into the predictive model of the content popularity, explicitly looking at the impact of influential users on their followers to persuade them to contribute to brand post popularity.

The main difference between these two models is that ETAS model is a Bayesian version of self-exciting process model with a dependent mark that treats the number of followers as a mark. Both formulations essentially lead to the same model. However, ETAS model leads us to a more accurate estimation of parameters due to the underlying Bayesian inference.

In Hassan Zadeh and Sharda (2014a), we observed that incorporating the number of followers into the predictive model of popularity of content presumably provides better results. In behavioral terms, it confirmed our hypothesis that the greater the number of followers per event, the greater the influence.

Also, in the follow-up paper (Hassan Zadeh and Sharda 2014b), we implemented the idea of excitation effects between different types of activities. To apply the mutual-exciting Hawkes model, we separated the dynamics of different types of activities that a given tweet receives over its lifetime in order to measure the popularity of a given online content. This type of Hawkes model includes the exciting effects among different types of users' activities on a brand's tweet into the predictive model. There were three point processes associated with each individual event type category (i.e. retweets, replies and favorites). It allowed us to capture the interacting effects between a stream of events from one to another. Our findings determined that incorporating the type of events into the predictive model of the brand post popularity provides a better understanding of such phenomena.

Several interesting managerial implications were derived from the mathematical models of Hawkes point process presented in previous research regarding the effects of different types (retweet, reply and favorite) of user activities on the popularity of the online content. For example, our model revealed that there are significant exciting effects between the same type of user activities as well as exciting effects between different types of user activities. Our model parameters indicated that retweeting is more likely to excite the other two types of events. This is consistent with the observed data as retweet action is more powerful. Furthermore, we concluded that it is more likely for users to behave like their friends and create the same type of event. For instance, once a brand's post receives "replies" multiple times, it can indicate origination of a conversation thread, and followers reply to the post rather than retweeting. Also, the mathematical model confirmed that users may use "mark as favorite" button to bookmark a tweet that contains subject matter they found interesting but they do not feel like broadcasting to the universe. By marking a tweet as favorite, users just take an action, have it saved and later refer back to it as needed.

Also, our model's parameters suggested that past retweets and replies visible in the timeline are more likely to excite a user to mark the tweet as favorite than being excited by itself. In behavioral terms, it is seldom that users check their friends' favorite tweets and decide to retweet, reply or favorite them.

As seen above, Hawkes process offers a powerful modeling method in social network analysis and has flexibility that can be applied to various kinds of time-dependent data to study events or behaviors of interest occurring in social media streams. There are several research opportunities in line with this research. Hawkes process-based analysis can be done in the context of Facebook, LinkedIn and other social media to study how the formation of relationships and interactions on Twitter is different than other social media platforms. Understanding which types of contents are appealing to audience and how users respond to various stimuli like videos, contests, applications or posts are something that can be more thoroughly investigated with the help of sentiment analysis tools. Also, many applications on social media involve more than a single type of event. It may be useful to treat repeated events of a single type (univariate) on multiple contents with multiple types as forming a multivariate point pattern.

In summary, recently there has been a growing interest to use Hawkes point process models for social network analysis. One of the convincing reasons for growing this interest is that Hawke process models offer a natural and traceable way of modeling time dependencies between events that are arisen as a result of branching, self-exciting and self-similarity behaviors in social networks. The underlying self- and mutual exciting mechanism in Hawkes process models is consistent with the structure observed in social networks. It leads to a nice representation that combines both branching process and conditional intensity representations in one solid model.

The reader should note that in this paper the terms "activity", "event" and "point" are used interchangeably. The term "activity" is used frequently in the context of online social network analysis; however the terms "event" and "point" are often used in the context of stochastic point processes.

5.6 Conclusion

This paper provides an introduction to the Hawkes point process as a very powerful and versatile tool for modeling and understanding social media traffic. These models are used to structure spatial and temporal dependencies between events that are arisen as a result of branching, self-exciting and self-similarity behaviors. It allows us to unravel implicit network structure that cannot fully be observed in online social networks, but have to be analyzed for different purposes in a natural, concrete and traceable computational way. Hawkes process models can be potentially applied to any kind of intensive time-stamped data to study events or behaviors of interest. Two of our previous research papers demonstrate the potential of the Hawkes processes in order to understand the dynamics of human interactions and collective behaviors on social media. Such analysis is fundamental to be able to predict how organization's social media campaign will evolve and grow to achieve the objectives of the campaign.

Biography

Amir Hassan Zadeh   is pursuing a Ph.D. in Management Science and Information Systems within the Spears School of Business at Oklahoma State University. He received his master's in Industrial and Systems Engineering from Amirkabir University of Technology, and his bachelor's from Department of Mathematics and Computer Science, Shahed University, Tehran, Iran. His research has been published in Decision Support Systems, Production Planning & Control, Advances in Intelligent and Soft Computing, African Journal of Business Management, and also conference proceedings of DSI, INFORMS and IEEE. His research addresses questions at the interface of operations management and information systems. His current research interest is also focused on the use of big data analytics for social media and recommender systems.

Ramesh Sharda   is the Vice Dean of the Watson Graduate School of Management, Watson/ConocoPhillips Chair and a Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University. He also serves as the Executive Director of the Ph.D. in Business for Executives Program. He has coauthored two textbooks (Business Intelligence and Analytics: Systems for Decision Support, 10th edition, Prentice Hall and Business Intelligence: A Managerial Perspective on Analytics, 3rd Edition, Prentice Hall). His research has been published in major journals in management science and information systems including Management Science, Operations Research, Information Systems Research, Decision Support Systems, Interfaces, INFORMS Journal on Computing, and many others. He is a member of the editorial boards of journals such as the Decision Support Systems and Information Systems Frontiers. He is currently serving as the Executive Director of Teradata University Network and received the 2013 INFORMS HG Computing Society Lifetime Service Award.

References

Adamopoulos, L. (1976). Cluster models for earthquakes: Regional comparisons. *Journal of the International Association for Mathematical Geology, 8*(4), 463–475.

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the second international symposium on information theory* (Vol. 1, pp. 610–624). New York: Springer.

Alexey, S., Martin, B. S., et al. (2011). Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems, 27*(11), 115013.

Bacry, E., Dayri, K., et al. (2012a). Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *The European Physical Journal B, 85*(5), 1–12.

Bacry, E., Delattre, S., et al. (2012b). Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance, 13*(1), 65–77.

Bauwens, L., & Hautsch, N. (2009). Modelling financial high frequency data using point processes. In T. Mikosch, J.-P. Kreiß, R. A. Davis, & T. G. Andersen (Eds.), *Handbook of financial time series* (pp. 953–979). Berlin/Heidelberg: Springer.

Blundell, C., Heller, K. A., Beck, J. M., & NIPS. (2012). Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems, 4*, 2600–2608.

Bowsher, C. (2003). *Modelling security market events in continuous time: Intensity based, multivariate point process models*. Oxford: Nuffield College.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics, 141*(2), 876–912.

Brillinger, D. R. (1975). The identification of point process systems. *The Annals of Probability, 3*(6), 909–924.

Brillinger, D. R. (1988). Maximum likelihood analysis of spike trains of interacting nerve cells. *Biological Cybernetics, 59*(3), 189–200.

Brillinger, D., Bryant, H., Jr., et al. (1976). Identification of synaptic interactions. *Biological Cybernetics, 22*(4), 213–228.

Cardanobile, S., & Rotter, S. (2010). Multiplicatively interacting point processes and applications to neural modeling. *Journal of Computational Neuroscience, 28*(2), 267–284.

Carlsson, J., Foo, M. C., Lee, H. H., & Shek, H. (2007). High frequency trade prediction with bivariate hawkes process. http://users.iems.northwestern.edu/~armbruster/2007msande444/report1b.pdf.

Chavez-Demoulin, V., & McGill, J. A. (2012). High-frequency financial data modeling using Hawkes processes. *Journal of Banking & Finance, 36*(12), 3415–3426.

Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences, 105*(41), 15649–15653.

Dahlhaus, R., Eichler, M., et al. (1997). Identification of synaptic connections in neural ensembles by graphical models. *Journal of Neuroscience Methods, 77*(1), 93–107.

Daley, D. J. (2006). *An introduction to the theory of point processes elementary theory and methods*. Retrieved from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=264777

Daley, D. J., & Vere-Jones, D. (2003). Conditional intensities and likelihoods. In *An introduction to the theory of point processes* (pp. 211–287). New York: Springer.

Daley, D. J., & Vere-Jones, D. (2003b). *An introduction to the theory of point processes* (Vol. 1).

Dassios, A., & Zhao, H. (2012). Ruin by dynamic contagion claims. *Insurance Mathematics and Economics, 51*(1), 93–106.

Engle, R. F., & Lunde, A. (2003). Trades and quotes: A bivariate point process. *Journal of Financial Econometrics, 1*(2), 159–188.

Lewis, E., Mohler, G., Brantingham, P. J., & Bertozzi, A. L. (2012). Self-exciting point process models of civilian deaths in Iraq. *Security Journal, 25*(3), 244–264.

Errais, E., Giesecke, K., et al. (2010). Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics, 1*(1), 642–665.

Giesecke, Kay., & Pascal, T. (2005). Dependent events and changes of time. Cornell University

Giesecke, K., Goldberg, L. R., et al. (2011). A top-down approach to multiname credit. *Operations Research, 59*(2), 283–300.

Golosovsky, M., & Solomon, S. (2012). Stochastic dynamical model of a growing citation network based on a self-exciting point process. *Physical Review Letters, 109*(9).

Guy, I., Jacovi, M., et al. (2010). Same places, same things, same people?: mining user similarity on social media. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 41–50). Savannah: ACM.

Halpin, P., & Boeck, P. (2013). Modelling dyadic interaction with Hawkes processes. *Psychometrika, 78*(4), 793–814.

Harte, D. (2010). PtProcess: An R package for modelling marked point processes indexed by time. *Journal of Statistical Software, 35*(8), 1–32.

Hassan Zadeh, A., & Sharda, R. (2014a). Modeling brand post popularity dynamics in online social networks. *Decision Support Systems, 65*(0), 59–68.

Hassan Zadeh, A., & Sharda, R. (2014b). *A point process framework for predicting popularity of online content in online social networks*. Available at SSRN 2331565.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika, 58*(1), 83–90.

Hawkes, A. G., & Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability, 11*(3), 493–503.

Hegemann, R., Lewis, E., et al. (2013). An "Estimate & Score Algorithm" for simultaneous parameter estimation and reconstruction of incomplete data on social networks. *Security Informatics, 2*(1), 1–13.

Herrera, R. (2013). Energy risk management through self-exciting marked point process. *Energy Economics, 38*, 64–76.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

Kim, H. (2011). *Spatio-temporal point process models for the spread of avian in anfuenza virus (H5N1)*. Retrieved from http://digitalassets.lib.berkeley.edu/etd/ucb/text/Kim_berkeley_0028E_11401.pdf

Krumin, M., Reutsky, I., & Shoham, S. (2010). Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input. *Frontiers in Computational Neuroscience, 4*, 147.

Lawrence, M., & Michael, E. C. (2010). Hawkes process as a model of social interactions: a view on video dynamics. *Journal of Physics A: Mathematical and Theoretical, 43*(4), 045101.

Lewis, E., Bertozzi, A. L., Mohler, G., & Brantingham, P. J. (2012). Self-exciting point process models of civilian deaths in Iraq. *Security Journal, 25*(3), 244–264.

Liniger, T. J. (2009). *Multivariate Hawkes processes*. Doctoral dissertation, Eidgenössische Technische Hochschule ETH Zürich, Nr. 18403.

Masuda, N., Takaguchi, T., Sato, N., & Yano, K. (2013). Self-exciting point process modeling of conversation event sequences. In *Temporal Networks* (pp. 245–264). Springer Berlin Heidelberg.

Meyer, S. (2009). *Spatio-temporal infectious disease epidemiology based on point processes*. Retrieved from http://epub.ub.uni-muenchen.de/11703/1/MA_Meyer.pdf

Mohler, G. (2013). Modeling and estimation of multi-source clustering in crime and security data. *The Annals of Applied Statistics, 7*(3), 1525–1539.

Mohler, G. O., Short, M. B., et al. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association, 106*(493), 100–108.

Møller, J., & Rasmussen, J. G. (2005). Perfect simulation of Hawkes processes. *Advances in Applied Probability, 37*(3), 629–646.

Ogata, Y. (1988). Statistical models for Earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association, 83*(401), 9–27.

Ogata, Y. (1999). Seismicity analysis through point-process modeling: A review. In M. Wyss, K. Shimazaki, & A. Ito (Eds.), *Seismicity patterns, their statistical significance and physical meaning* (pp. 471–507). Basel: Birkhäuser.

Ogata, Y. (2006). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory, 27*(1), 23–31.

Ogata, Y., & Vere-Jones, D. (1984). Inference for earthquake models: A self-correcting model. *Stochastic Processes and their Applications, 17*(2), 337–347.

Paraschiv-Ionescu, A., Buchser, E., & Aminian, K. (2013). Unraveling dynamics of human physical activity patterns in chronic pain conditions. *Scientific Reports, 3*, 2019. doi:10.1038/srep02019.

Pernice, V., Staude, B., et al. (2011). How structure determines correlations in neuronal networks. *PLoS Computational Biology, 7*(5), e1002059.

Porter, M. D., & White, G. (2012). Self-exciting hurdle models for terrorist activity. *Annals of Applied Statistics, 6*(1), 106–124.

Rasmussen, J. (2013). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability, 15*(3), 623–642.

Rathbun, S. L., Shiffman, S., & Gwaltney, C. J. (2006). Point process models for event history data: applications in behavioral science. *Models for intensive longitudinal data, 219*.

Reynaud-Bouret, P., Rivoirard, V. et al. (2013). *Inference of functional connectivity in neurosciences via Hawkes processes*. Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE, Piscataway.

Schoenberg, F. P., Cochran, J. J., et al. (2010). Introduction to point processes. In *Wiley encyclopedia of operations research and management science*. John Wiley & Sons, Inc.

Steeg, G. V., & Galstyan, A. (2012). Information transfer in social media. In *Proceedings of the 21st international conference on World Wide Web* (pp. 509–518). Lyon: ACM.

Veen, A., & Schoenberg, F. P. (2008). Estimation of space–Time branching process models in seismology using an EM–Type algorithm. *Journal of the American Statistical Association, 103*(482), 614–624.

Wang, T., Bebbington, M., et al. (2012). Markov-modulated Hawkes process with stepwise decay. *Annals of the Institute of Statistical Mathematics, 64*(3), 521–544.

Xu, L., Duan, J. A., Whinston, A. B. (2014). *Path to purchase: A mutually exciting point process model for online advertising and conversion. Management Science, 60*(6), 1392–1412.

Yang, S. H., & Zha, H. (2013). Mixture of mutually exciting processes for viral diffusion. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (pp. 1–9).

Zammit-Mangion, A., Dewar, M., et al. (2012). Point process modelling of the Afghan War Diary. *Proceedings of the National Academy of Sciences, 109*(31), 12414–12419.

Zheng, B., Roueff, F., et al. (2014). Modelling bid and ask prices using constrained Hawkes processes: Ergodicity and scaling limit. *SIAM Journal on Financial Mathematics, 5*(1), 99–136.

Chapter 6
Using Academic Analytics to Predict Dropout Risk in E-Learning Courses

Rajeev Bukralia, Amit V. Deokar, and Surendra Sarnikar

Abstract Information technology is reshaping higher education globally and analytics can help provide insights into complex issues in higher education, such as student recruitment, enrollment, retention, student learning, and graduation. Student retention, in particular, is a major issue in higher education, since it has an impact on students, institutions, and society. With the rapid growth in online enrollment, coupled with a higher dropout rate, more students are at risk of dropping out of online courses. Early identification of students who are at risk to drop out is imperative for preventing student dropout. This study develops a model to predict real-time dropout risk for each student while an online course is being taught. The model developed in this research utilizes a combination of variables from the Student Information Systems (SIS) and Course Management System (CMS). SIS data consists of ten independent variables, which provide a baseline risk score for each student at the beginning of the course. CMS data consists of seven independent variables that provide a dynamic risk score as the course progresses. Furthermore, the study provides an evaluation of various data mining techniques for their predictive accuracy and performance to build the predictive model and risk scores. Based on predictive model, the study presents a recommender system framework, to generate alerts and recommendations for students, instructors, and staff to facilitate early and effective intervention. The study results show that the boosted C5.0 decision tree model achieves 90.97 % overall predictive accuracy in predicting student dropout in online courses.

R. Bukralia (✉)
Information & Computing Science, University of Wisconsin – Green Bay,
2420 Nicolet Drive, MAC C336, Green Bay, WI 54311-7001, USA
e-mail: bukralir@uwgb.edu

A.V. Deokar
The Pennsylvania State University, 5101 Jordan Road, 268 Burke Center,
Erie, PA 16563, USA
e-mail: amit.deokar@psu.edu

S. Sarnikar
College of Business and Information Systems, Dakota State University,
820 N. Washington Ave., East Hall 7, Madison, SD 57042, USA
e-mail: surendra.sarnikar@dsu.edu

6.1   Introduction

Information technology is reshaping higher education globally and analytics can help provide insights into complex issues in higher education, such as student recruitment, enrollment, retention, student learning, and graduation. Student retention, in particular, has become a major issue in higher education, since it has an impact on students, institutions, and society. This study deals with the important issue of student retention at a micro level, i.e. at the course level, since course completion is considered to be "the smallest unit of analysis with respect to retention" (Hegedorn 2005). Student retention is a complex and widespread issue – 40 % of students leave higher education without getting a college degree (Porter 1989). Both the student and the institution need to shoulder the responsibility of ensuring that students can succeed in higher education. Therefore, institutions of higher education are looking for strategies to improve their retention rate. It is important to identify at-risk students – or those students who have a greater likelihood of dropping out of a course or program – as this can allow instructors and advisors to proactively implement appropriate retention strategies. Studies have shown much higher dropout rates in online courses compared to face-to-face courses. Several studies have shown that the dropout rate is 10–40 % higher in online courses compared to traditional, face-to-face courses (Carr 2000; Diaz 2000; Lynch 2001).

Online course dropout needs to be addressed to improve institutional effectiveness and student success. One of the key elements in reducing the e-learning dropout rate is the accurate and prompt identification of students who may be at greater risk to drop out (Lykourentzou et al. 2009). Previous studies have focused on identifying students who are more likely to drop out using academic and demographic data, obtained from Student Information Systems (SIS) (Kiser and Price 2007; Park 2007). Online courses are generally taught using a Course Management System (CMS), which can provide detailed data about student activity in the course. There is a need to develop models that can predict real-time dropout risk for each student while an online course is being taught. Using both SIS and CMS data, a predictive model can provide a more accurate, real-time dropout risk for each student while the course is in progress.

The model developed in this research study utilizes a combination of variables from the SIS to provide a baseline risk score for each student at the beginning of the course. Data from the CMS is used, in combination with the baseline prediction, to provide a dynamic risk score as the course progresses. This study identifies and analyzes various SIS-based and CMS-based variables to predict dropout risk for students in online courses and evaluates various data mining techniques for their predictive accuracy and performance to build the predictive model and risk scores.

The model leverages both SIS (historical) and CMS (time-variant) data to improve on the predictive accuracy. The model provides a basis for building early alert and recommender systems so instructors and retention personnel can deploy proactive strategies before an online student drops out.

The rest of the paper includes the literature review in Sect. 6.2. Section 6.3 contains the methodology of the study. Section 6.4 presents the data analysis and results, and the proposed recommender system framework, followed by Conclusion.

6.2   Literature Review

6.2.1   Factors Affecting Student Retention and Dropout

Early detection of at-risk students and appropriate intervention strategies are a key in retention. Seidman (2005) developed a formula of student retention: *Retention = Early Identification + (Early + Intensive + Continuous) Intervention*. The formula emphasizes the role of early identification and intervention in improving student retention. Tinto (1987) introduced the importance of student integration (both socially and academically) in the prediction of student retention. Tinto (1993) reported that the factors in students dropping out include academic difficulty, adjustment problems, lack of clear academic and career goals, and poor social integration with the college community.

There are studies that have investigated the role of academic and non-academic variables in student retention in face-to-face programs, such as high school GPA and ACT scores – both of which were found to be good predictors of retention rates in face-to-face programs (Campbell and Oblinger 2007; Lotkowski et al. 2004). Lotkowski et al. (2004) note that college performance was strongly related to ACT scores, high school GPA (HSGPA), and socio-economic status (SES), as well as academic self-confidence and achievement motivation. Their study identified a strong relationship between all these factors and college performance. Some studies have shown that retention is influenced by college GPA once a student is in college (Cabrera et al. 1993; Mangold et al. 2002; O'Brien and Shedd 2001).

Non-academic factors, typically assessed once the student is enrolled, can also affect retention (Braxton 2000). Non-academic factors that have been known to influence retention include level of commitment to obtaining a degree, academic self-confidence, academic skills, time management skills, study skills, study habits, and level of academic and social integration into the institution (Lotkowski et al. 2004).

Kiser and Price (2007) examined the predictive accuracy of high school GPA (HSGPA), resident's location, cumulative hours taken, mother's education level, father's education level, and gender on persistence of college freshmen to sophomore year (freshman retention). Their model used logistic regression on a dataset of 1,014 students and found that cumulative hours taken was statistically significant for the overall model. Cumulative credit hours have financial implications for

students; as the number of hours grows, student investment grows. According to Parker and Greenlee (1997), financial problems, family complications, work schedule conflicts, and poor academic performance (in the order of importance) were the most important factors in persistence of nontraditional students. Bean and Metzner (1985) identified four factors that affect persistence of students, especially nontraditional students namely: academic variables, background and defining variables, environmental variables, academic and psychological outcomes.

Dutton and Perry (2002) examined the characteristics of students enrolled in online courses and how those students differed from their peers in traditional face-to-face courses. The study found that online students were older, more likely to have job responsibilities, and required more flexibility for their studies. In terms of the student gender, Rovai (2003) and Whiteman (2004) found that females tend to be more successful at online courses than males. However, another study showed that gender had no correlation with persistence in an e-learning program (Kemp 2002). Some studies have pointed out the relevance of age as a predictor of dropout rates in e-learning (Muse 2003; Whiteman 2004). Diaz, though, found that online students were older than traditional face-to-face students, but there was not a significant correlation between age and retention (Diaz 2000). A more recent study conducted by Park (2007), on a large population of about 89,000 students enrolled in online courses showed that age, gender, ethnicity, and financial aid eligibility were good predictors of successful course completion. Furthermore, Yu et al (2007) reported that earned credit hours were linked with student retention in online courses; the study also showed a correlation between the location of the student – in-state or out-of-state – and retention.

The usage data from the CMS can provide useful insights about students' learning behavior. The data on student online activity in the CMS can provide an early indicator of student academic performance (Wang and Newlin 2002). CMS data can provide useful information about study patterns, engagement, and participation in an online course (Dawson et al. 2008). It has been argued that institutional CMS data can offer new insights into student success, and help identify students who are at risk of dropout or course failure (Campbell et al. 2007). Furthermore, Campbell et al. (2006) found that student SAT scores were mildly predictive for future student success; however, when a second variable, CMS login, was added, it tripled the predictive accuracy of the model.

Studies have used CMS activity logs to analyze learner paths and learning behavioral patterns (Bellaachia et al. 2006; Hung and Zhang 2008), to elicit the motivational level of the students towards the course (Cocea and Weibelzahl 2007), and to assess the performance of learners (Chen et al. 2007). Studies have indicated a significant relationship exists between CMS variables and academic performance (Macfadyen and Dawson 2010; Morris et al. 2005). The results indicate that engaged students are more likely to successfully complete a course than students who are less interactive and less involved with their peers. Research studies have been conducted to mine CMS data (such as number and duration of online sessions, discussion messages read or posted, content modules or pages visited, and number of quizzes or assignments completed or submitted) to identify students who are more

likely to drop out of a course or receive a failing grade. CMS datasets can be captured in real time and can be mined to provide information about how a student is progressing in the course (Macfadyen and Dawson 2010).

6.2.2 Data Analysis and Mining Techniques in Dropout Prediction

There are several analysis and mining methods used to analyze and mine student data. Logistic regression (LR) is heavily used in predictive models that have a binary dependent variable (response variable). Logistic regression has been widely used in business to predict customer attrition events, sales events for a product, or any event that has a binary outcome (Nisbet et al. 2009). Many studies that were conducted to identify high-risk students used statistical models based on logistic regression (Araque et al. 2009; Newell 2007; Willging and Johnson 2004). Logistic regression and survival analysis were used to build a competing risk model of retention (Denson and Schumacker 1996). The rationale of using logistic regression in the retention problem is that outcome is typically binary (enrolled or not enrolled) and because probability estimates can be calculated for combinations of multiple independent variables (Pittman 2008).

Park and Choi (2009) used logistic regression analysis to determine how well four variables (family support, organizational support, satisfaction, and relevance) predicted learners' decisions to drop out. Furthermore, Roblyer et al. (2008) gathered data on student characteristics regarding dropout in an online environment. They analyzed a dataset of over 2,100 students using binary logistic regression, with an overall correct classification rate of 79.3 %. Logistic regression was also used by Allen and Robbins (2008) to predict persistence over a large dataset of 50,000 students. The study used three variables: students' vocational interest, their academic preparation, and their first-year academic performance. The study concluded that prior academic performance was the critical element in predicting persistence. A logistic regression model was used to identify at-risk students using CMS variables such as number of messages posted, assignments completed, and total time spent. This model accurately classified at-risk students with 73.7 % accuracy (Macfadyen and Dawson 2010).

On the other hand, several machine learning techniques, such as decision trees, neural networks, Support Vector Machines (SVM), and naïve Bayes are appropriate for binary classification. In that regard, Etchells et al. (2006) used neural networks for predicting students' final grades. Furthermore, Herzog (2006) and Campbell (2008) compared neural networks with regression to estimate student retention and degree-completion time. One of the most commonly used techniques of data mining is a decision tree, which is a technique used in solving classification problems. In this context, Breiman et al. (1984) used a type of decision tree at the University of California San Diego Medical Center to identify high-risk patients. Furthermore, Cocea and Weibelzahl (2007) used a decision tree to analyze log files from the CMS

to see the relationship between time spent reading and student motivation. Muehlenbrock (2005) applied a C4.5 decision tree model to analyze user actions in the CMS to predict future uses of the learning environment.

Given the rich information provided by the CMS, it is argued that CMS data, in addition to SIS data, may provide a more accurate, real-time dropout prediction at the course level. In this study, we present an approach that utilizes a combination of SIS variables to provide a baseline prediction for each student at the beginning of the course. CMS data from the current course is used, along with the baseline prediction, to provide a real-time risk score as the course progresses. The purpose of dynamic prediction by adjusting the baseline prediction with CMS log activity on a regular basis (daily or weekly) is to provide a more accurate prediction of likely dropouts. Up to our knowledge, no study has been conducted that has utilized both SIS (static data) and CMS (time-variant) data together to make a dynamic prediction while the course is in progress.

6.3   Research Approach

In this section we discuss the overall research process including the identification of relevant constructs and predictor variables and the evaluation criteria. An overview of the overall research process is shown in Fig. 6.1. We approach the problem of prediction whether a student will complete or drop a course as a binary classification problem. First, we begin with a literature review focused on retention theories and dropout studies identifying constructs and variables important to course completion or dropout. The following constructs were identified that were found to be useful for predicting online course dropout in the literature review: academic ability, financial support, academic goals, technology preparedness, demographics, course motivation and engagement, and course characteristics. The constructs were mapped to appropriate variables. The variables were selected based on the literature review, as well as available resources and availability of data. In the context of SIS variables, data was collected from the Student Information System.

The literature review helped to identify various data mining techniques used in binary classification, in addition to hybrid machine learning techniques proposed in this study. SIS data was analyzed using descriptive statistics and various data mining techniques (LR, ANN, SVM, DT, NB, GA-ANN, and GA-NB). Using the accuracy metrics, each technique is evaluated. The most accurate technique (Algorithm I) is used to build the SIS Predictive Model, which provides a baseline dropout risk score for each student. CMS data is gathered from the Course Management System at the end of the third and seventh day from the beginning of the course. The data reflects the changes in student usage of the CMS in each online course. The data is analyzed using Algorithm II, which suits the time-sensitive nature of CMS data, since the usage of the CMS is updated daily. The dynamic prediction (CMS Predictive Model) provides a real-time dropout risk score, which can be updated as frequently as CMS data become updated.
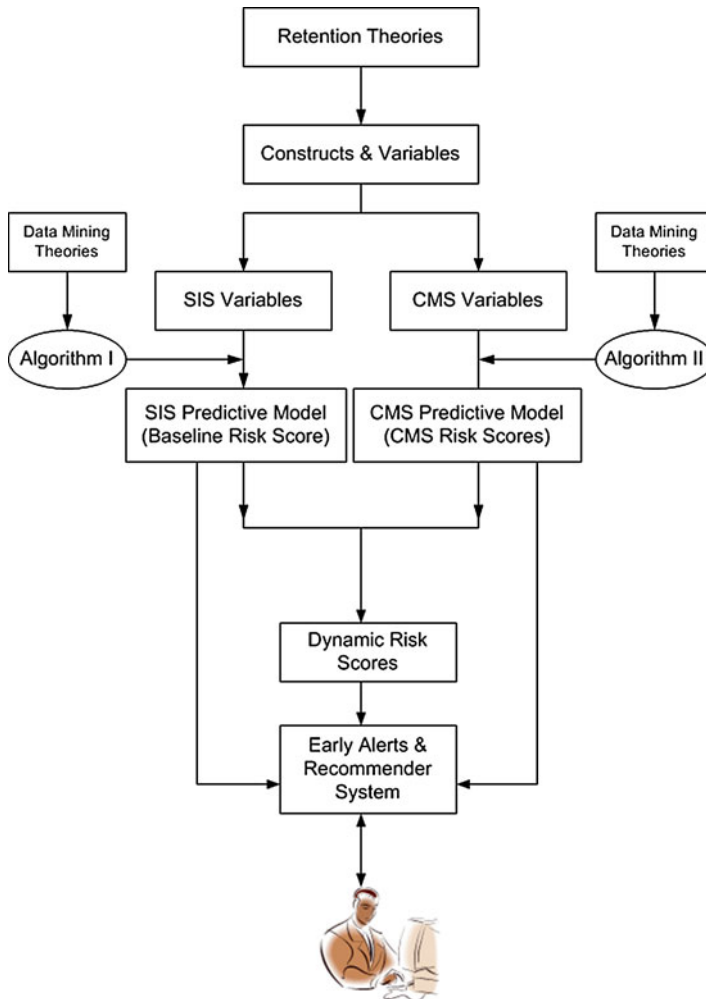
Fig. 6.1 Research approach

The recommender system uses the predictive models to identify students who are at greater risk of dropping out of the course. The purpose of the recommender system is twofold: to create and send alerts to retention staff, students, and instructors at a predefined time and predefined score; and to recommend customized interventions to at-risk students. For example, if a student is given a high risk score because of financial aid status, the student will be directed to work with financial aid staff. Similarly, if a student has a high risk score due to not having taken an online course previously, then the student will be directed to online training videos relevant for e-learning. Figure 6.1 summarizes the key steps of the research approach.

6.3.1 Constructs and Variables

The following independent constructs were identified for this study to predict dropout: academic ability, financial support, academic goals, technology preparedness, demographics, course motivation and engagement, and course characteristics. Each identified construct was further mapped into appropriate variables. For instance, the dependent construct is the course completion, which includes course completion status variable. Regarding the independent variables, the study identifies the following variables for each construct grounded in the literature:

1. Academic ability: ACT score (Lotkowski et al. 2004; Vare et al 2000), high school GPA (Vare et al. 2000), current college GPA (Diaz 2000).
2. Financial support: financial aid status (Newell 2007).
3. Academic goals: credit hours completed (Yu et al. 2007), previous drops, degree seeking status (Bukralia et al. 2009).
4. Technology preparedness: previous online course completion (Osborn 2001).
5. Demographics : gender (Kemp 2002; Ross and Powell 1990; Whiteman 2004) and age (Diaz 2000; Muse 2003; Whiteman 2004).
6. Course engagement and motivation: total number of logins in CMS by the student, total time spent in CMS, and Number of days since last login in CMS.
7. Course characteristics: credit hours, course level, and course prefix.

6.3.2 Evaluation Criteria

Since superfluous variables can lead to inaccuracies, it is important to remove them. Stepwise selection (in regression) and feature selection (in machine learning) are common approaches for removing superfluous variables. To build a binary classification based predictive model, the following questions should be carefully addressed: (a) Which classification algorithms are appropriate in relation to problem characteristics? (b) How is the performance of selected classification algorithms evaluated?

The data mining techniques are selected based on the nature of the prediction (whether both class labels and probability of class membership are needed), nature of independent variables (continuous, categorical, or both), and the algorithmic approach (black box or white box). Both statistical and machine learning algorithms are used to identify class membership (for example, whether a student completed or dropped the course). A statistical technique (logistic regression), machine learning techniques (ANN, DT, SVM, and NB), and hybrid machine learning techniques (or ensemble methods such as boosted trees) are used to compare their predictive accuracy. The classification models are evaluated using two criteria: discrimination and calibration. Discrimination is used to demonstrate how well two classes (0 or 1) are separated, while calibration provides the accuracy of probability estimates.

Common measures of discrimination include accuracy, precision, specificity, and sensitivity where:

$$\text{Sensitivity} = \text{TP} / \text{P} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN} / \text{N} = \text{TN} / (\text{FP} + \text{TN})$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{P} + \text{N})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}).$$

6.4   Data Analysis and Predictive Modeling

This section discusses the step-by-step process for collecting, preparing, and analyzing both SIS and CMS datasets. The process starts with the collection and preparation of SIS data for data analysis. The SIS datasets are analyzed using descriptive statistics and various data mining techniques to compare the accuracy in terms of identifying dropout cases. Similarly, CMS data are collected, cleansed, and analyzed. Various data mining techniques are evaluated for accuracy and performance with CMS datasets. Using the most accurate technique, a baseline risk scoring model is built using SIS data and a dynamic risk scoring model is built using both SIS and CMS data. The section concludes with the evaluation of the baseline and dynamic risk models to identify at-risk students.

6.4.1   SIS Data Analysis

Table 6.1 presents the variables that were chosen to be used in the SIS dataset. The dataset contains 1,376 student records from online courses offered in the Fall 2009 semester, and was extracted from the SIS (Datatel Colleague System: http://www.datatel.com/products/products_a-z/colleague_student_fa.cfm).

Table 6.1   Variables in the Student Information System (SIS) dataset

No.	Variable	Data type	No.	Variable	Data type
1	ACT score	Numeric	7	Degree seeking status	Binary nominal
2	High school GPA	Numeric	8	Previous online course completion	Binary nominal
3	Current college GPA	Numeric	9	Gender	Binary nominal
4	Financial aid status	Binary nominal	10	Age	Numeric
5	Credit hours completed	Numeric	11	Course status (dependent variable)	Binary nominal
6	Previous drops (course withdrawals)	Binary nominal			

An examination of the dataset showed that it had some noisy data, missing values, and outliers. Many of the noisy data issues were related to human error in data entry. 1,113 records were from students who completed an online course, with 263 records from students who dropped out of an online course with 19.11 % (263 out of 1,376 total records) of students dropped their online course. To describe data, descriptive statistics were performed using WEKA and SPSS statistics packages. The analysis revealed the following dataset characteristics:

1. 47.82 % of students did not have financial aid; while 52.18 % of students received financial aid.
2. College GPA had approximately 19 % missing values, with a mean of 2.5 and a standard deviation of 1.31.
3. High school GPA had 53 % missing values, with a mean of 3.18 and a standard deviation of .58 – which was understandable, as high school GPAs are above 2.0, since students who had a GPA in their records had completed high school before college admission.
4. The histograms show that the dataset had a higher number of records for female students than for males (974 records compared to 311 records).

In order to get accurate data analysis, it is essential to prepare the dataset using data preprocessing and formatting so that appropriate data mining techniques can be used to analyze data. As the dataset is not large, the deletion of records was not found to be a good strategy for this dataset, as it would have reduced the number of records to about merely 300. Records with outliers were removed and missing data were replaced using the mean value for numerical variables (ACT score, age, and college GPA). Missing data for categorical variables (such as degree seeking status) were set as zero (non-degree seeking, no financial aid, no previous online course, and female gender). The initial dataset was imbalanced, as the dependent variable (course status) had more records for students completing the course (course status=1) than for students who dropped out (course status=0).

Using WEKA, the following data mining techniques were performed on the dataset using 10fold cross validation to create a predictive model: Decision Trees using J.48 algorithm (DT), Naïve Bayes (NB), Logistic Regression (LR), Artificial Neural Networks with Multilayer Perceptron algorithm (ANN), and Support Vector Machines (SVM). Table 6.2 shows the coincidence matrix for the comparative accuracy of each data mining technique using the unbalanced SIS dataset. Coincidence Matrix also known as a misclassification matrix, provides information about true positives, true negatives, false positives, and false negatives.

As Table 6.2 shows, each technique shows an overall accuracy of around 80 %. The data analysis showed that ANN (MLP) technique provided the best predictive accuracy for students who completed the course (88.01 %) and for students who dropped out (55.55 %). ANN (MLP) also had the best overall predictive accuracy among tested techniques. With this dataset, true negatives (students who actually dropped out) were low. Since the focus of this study was to predict students who are likely to dropout, this accuracy was concerning. The accuracy of predicting true negatives for DT, NB, LR, ANN (MLP), and SVM was 0 %, 21.73 %, 31.57 %,

Table 6.2 Coincidence matrix for various techniques using unbalanced SIS dataset. n = 1,376

		Dropped (0)	Completed (1)	Sum	Per class accuracy (%)	Overall accuracy (%)	ROC area
DT (J.48)	Dropped (0)	0	0	0	0	80.88	0.495
	Completed (1)	263	1,113	1,376	80.88		
Naïve Bayes	Dropped (0)	5	18	23	21.73	79.94	0.585
	Completed (1)	258	1,095	1,353	80.93		
LR	Dropped (0)	6	13	19	31.57	80.37	0.632
	Completed (1)	257	1,100	1,357	81.06		
ANN (MLP)	Dropped (0)	125	100	225	55.55	82.70	0.73
	Completed (1)	138	1,013	1,151	88.01		
SVM	Dropped (0)	0	2	2	0	80.74	0.499
	Completed (1)	263	1,111	1,374	80.85		

Table 6.3 Coincidence matrix for various techniques using balanced SIS dataset. n = 525

		Dropped (0)	Completed (1)	Sum	Per class accuracy (%)	Overall accuracy (%)	ROC area
DT (J.48)	Dropped (0)	177	53	230	76.95	73.52	0.739
	Completed (1)	86	209	295	70.84		
Naïve Bayes	Dropped (0)	165	71	236	69.10	67.80	0.759
	Completed (1)	98	191	289	66.08		
LR	Dropped (0)	183	71	254	72.04	71.23	0.786
	Completed (1)	80	191	271	70.47		
ANN (MLP)	Dropped (0)	184	64	248	74.19	72.76	0.778
	Completed (1)	79	198	277	71.48		
SVM	Dropped (0)	173	60	233	74.24	71.42	0.714
	Completed (1)	90	202	292	69.17		

55.55 %, and 0 %, respectively. DT and SVM were especially poor in identifying true negatives. ANN (MLP) worked best to find true negatives (55.55 %); however that percentage was not acceptable either.

In order to correct the bias in the dataset, most duplicated student records were removed. If a student registered for more than one online course and completed one or more courses and dropped one or more courses, then one record for each course was kept to remove bias. The removal of duplicated student records provided a balanced but much smaller dataset. The new dataset included an almost equal number of records of students who completed or dropped out of a course. Tables 6.3 and 6.4 shows the coincidence matrix and the evaluation measures for each data mining technique using the balanced dataset of 525 student records.

The overall predictive accuracy with the balanced dataset was slightly less than the accuracy with the unbalanced dataset. This was expected, because the unbalanced dataset provided an overly optimistic prediction for students who completed the

Table 6.4 Evaluation measures for various techniques using balanced SIS dataset. n=525

	Root mean square error	Weighted TP rate	Weighted FP rate	Weighted precision	Weighted recall
DT (J.48)	0.445	0.735	0.265	0.73	0.73
Naïve Bayes	0.488	0.678	0.322	0.68	0.68
LR	0.432	0.712	0.288	0.71	0.71
ANN (MLP)	0.469	0.728	0.272	0.72	0.72
SVM	0.534	0.714	0.285	0.71	0.71

course since it had more records for them in the dataset. The coincidence matrix shows that all techniques had a better accuracy of predicting dropout students from the balanced dataset (a predictive accuracy for dropout ranging from 69.1 % for naïve Bayes to 76.95 % for decision trees). In terms of predictive accuracy, decision trees performed better than other models.

Based on this observation, decision trees algorithms were further explored. It can be noted that in addition to the J.48 algorithm, C5.0 algorithm is also widely used in classification problems. C5.0 is a sophisticated decision tree algorithm, which works by splitting the sample based on the field that provides the maximum information gain. In this algorithm, each subsample is split multiple times based on various fields, and the process repeats until the subsamples cannot be split any further. The algorithm examines splits and prunes those that do not contribute significantly to the value of the model. It has been shown to work well with datasets that have missing values. Although the test/training data was cleaned, SIS data in production would contain missing values (such as the absence of HS GPA or ACT score). Other benefits of C5.0 are that it does not require long training times to estimate and it offers a powerful boosting method to increase the accuracy of classification. Accordingly, this study chose to compare the C5.0 algorithm with J.48 for decision trees.

In order to achieve better predictive accuracy compared to the accuracy of individual techniques, boosting was used. Boosting is an ensemble data mining technique. Ensemble data mining techniques leverage the power of multiple models, which consists of a set of individually trained classifiers (such as decision trees) whose predictions are combined when classifying novel instances. Ensemble techniques have been found to improve the predictive accuracy of the classifier (Opitz and Maclin 1999). To create a boosted C5.0 model a set of related models are created. Boosting in C5.0 builds multiple models in a sequence. The first model is built in the usual C5.0 tree. The second model is built using the records that were misclassified by the first model. Then a third model is built using the errors of the second model, and so on. Finally, cases are classified by applying the whole set of models to them, using a weighted voting procedure to combine the separate predictions into one overall prediction.

The balanced dataset containing 525 student records was analyzed with C5.0 algorithm in IBM Modeler software (WEKA does not provide this algorithm) using 10fold cross validation. Table 6.5 shows the comparative coincidence matrix and the ROC area for J.48, C5.0, and boosted C5.0 decision trees.

Table 6.5 Coincidence matrix for decision tree techniques for balanced SIS dataset. n = 525

		Dropped (0)	Completed (1)	Sum	Per class accuracy (%)	Overall accuracy (%)	ROC area
DT (J.48)	**Dropped (0)**	177	53	230	76.95	73.52	0.739
	Completed (1)	86	209	295	70.84		
DT (C5.0 without Boosting)	**Dropped (0)**	218	27	245	88.97	86.29	0.886
	Completed (1)	45	235	280	83.92		
DT (C5.0 Boosted)	**Dropped (0)**	229	15	244	93.85	90.67	0.965
	Completed (1)	34	247	281	87.90		

The analysis of the SIS dataset with C5.0 decision trees (with and without boosting) provided a greater accuracy than J.48 decision trees of predicting both types of cases – students who dropped the course and students who completed the course. Compared to 73.52 % accuracy for J.48, C5.0 provided 86.29 % accuracy; while boosted C5.0 decision trees were able to identify 90.67 % of students accurately. Since the purpose of the model is to accurately identify students who are at greater risk of dropping out, it is important to examine the accuracy of true negatives. The boosted C5.0 decision trees model accurately identified 93.85 % of students who dropped out, compared to 88.97 % without boosting, and 76.95 % for J.48. Figure 6.2 shows the boosted C5.0 decision tree for the SIS dataset. As the figure shows, the tree is initially split for the credit hours (college credit hours taken) node. This node indicates the split for credit hours equal to or less than 68, or for greater than 68 h. Furthermore, college GPA is equal to or less than 3.563, with age equal to or less than 22, then the student would drop out.

The boosted C5.0 decision tree model provided the relative importance of predictors, or independent variables. No predictor was found to be especially strong; however college credit hours (Credit Hours) and Age were found to be slightly more important than the other variables. The gain chart (Fig. 6.3) for the model suggests that the model can identify close to 100 % of cases with only 60 % of the total selections. In a well-fitted model, the response bow is well arched in the middle of the file. The gain chart of the boosted C5.0 model has a well arched bow in the middle, which reflects that the model is well-fitted and has long term stability.

Based on the comparative accuracy and evaluation measures, the boosted C5.0 decision tree model was used to build a baseline risk score for students. The boosted C5.0 model predicted the most likely value of course status (0 or 1). This value in the model is represented as $C-Current_Status (predicted course status). The model also provides the propensity scores, represented as $CC-Current_Status, and adjusted propensity scores, represented as $CRP-Current_Status. Therefore, Baseline Risk Score is computed by the following equation:

```
Baseline Risk Score = 100 - (Adjusted Propensity Score * 100)
Baseline Risk Score = 100 - ($CRP * 100)
```

```
⊟ Credit_Hours <= 68 [ Mode: 0 ]
  ⊟ GPA <= 3.563 [ Mode: 0 ]
    ⊟ Age <= 43 [ Mode: 0 ]
      ⊟ DegSkg = 1.000 [ Mode: 0 ]
          Age <= 22 [ Mode: 0 ]  ⇨  0.0
        ⊟ Age > 22 [ Mode: 0 ]
          ⊟ ACTComp <= 20 [ Mode: 1 ]
              Credit_Hours <= 58 [ Mode: 1 ]  ⇨  1.0
              Credit_Hours > 58 [ Mode: 0 ]  ⇨  0.0
          ⊟ ACTComp > 20 [ Mode: 0 ]
              ACTComp <= 21 [ Mode: 0 ]  ⇨  0.0
            ⊟ ACTComp > 21 [ Mode: 1 ]
                ACTComp <= 25 [ Mode: 1 ]  ⇨  1.0
                ACTComp > 25 [ Mode: 0 ]  ⇨  0.0
          DegSkg = 0.000 [ Mode: 0 ]  ⇨  0.0
      ⊟ Age > 43 [ Mode: 1 ]
        ⊟ Past_W = 1.000 [ Mode: 0 ]
            Age <= 47 [ Mode: 1 ]  ⇨  1.0
            Age > 47 [ Mode: 0 ]  ⇨  0.0
        ⊟ Past_W = 0.000 [ Mode: 1 ]
            Gender = 1.000 [ Mode: 0 ]  ⇨  0.0
            Gender = 0.000 [ Mode: 1 ]  ⇨  1.0
  ⊟ GPA > 3.563 [ Mode: 1 ]
    ⊟ Past_Online = 1.000 [ Mode: 1 ]
        Credit_Hours <= 51 [ Mode: 1 ]  ⇨  1.0
        Credit_Hours > 51 [ Mode: 0 ]  ⇨  0.0
      Past_Online = 0.000 [ Mode: 0 ]  ⇨  0.0
⊟ Credit_Hours > 68 [ Mode: 1 ]
    Age <= 22 [ Mode: 0 ]  ⇨  0.0
  ⊟ Age > 22 [ Mode: 1 ]
    ⊟ Past_W = 1.000 [ Mode: 1 ]
        Gender = 1.000 [ Mode: 1 ]  ⇨  1.0
      ⊟ Gender = 0.000 [ Mode: 1 ]
        ⊟ DegSkg = 1.000 [ Mode: 1 ]
          ⊟ Past_Online = 1.000 [ Mode: 1 ]
            ⊟ Credit_Hours <= 145 [ Mode: 1 ]
              ⊟ FinancialAid = 1.000 [ Mode: 1 ]
                ⊟ HS_GPA <= 2.310 [ Mode: 0 ]
                    HS_GPA <= 1.987 [ Mode: 1 ]  ⇨  1.0
                    HS_GPA > 1.987 [ Mode: 0 ]  ⇨  0.0
                  HS_GPA > 2.310 [ Mode: 1 ]  ⇨  1.0
              ⊟ FinancialAid = 0.000 [ Mode: 1 ]
                  GPA <= 2.625 [ Mode: 0 ]  ⇨  0.0
                ⊟ GPA > 2.625 [ Mode: 1 ]
                    Age <= 24 [ Mode: 0 ]  ⇨  0.0
                    Age > 24 [ Mode: 1 ]  ⇨  1.0
              Credit_Hours > 145 [ Mode: 1 ]  ⇨  1.0
            Past_Online = 0.000 [ Mode: 1 ]  ⇨  1.0
          DegSkg = 0.000 [ Mode: 1 ]  ⇨  1.0
    ⊟ Past_W = 0.000 [ Mode: 1 ]
      ⊟ Credit_Hours <= 107 [ Mode: 0 ]
          ACTComp <= 20 [ Mode: 1 ]  ⇨  1.0
        ⊟ ACTComp > 20 [ Mode: 0 ]
            GPA <= 3.513 [ Mode: 0 ]  ⇨  0.0
```
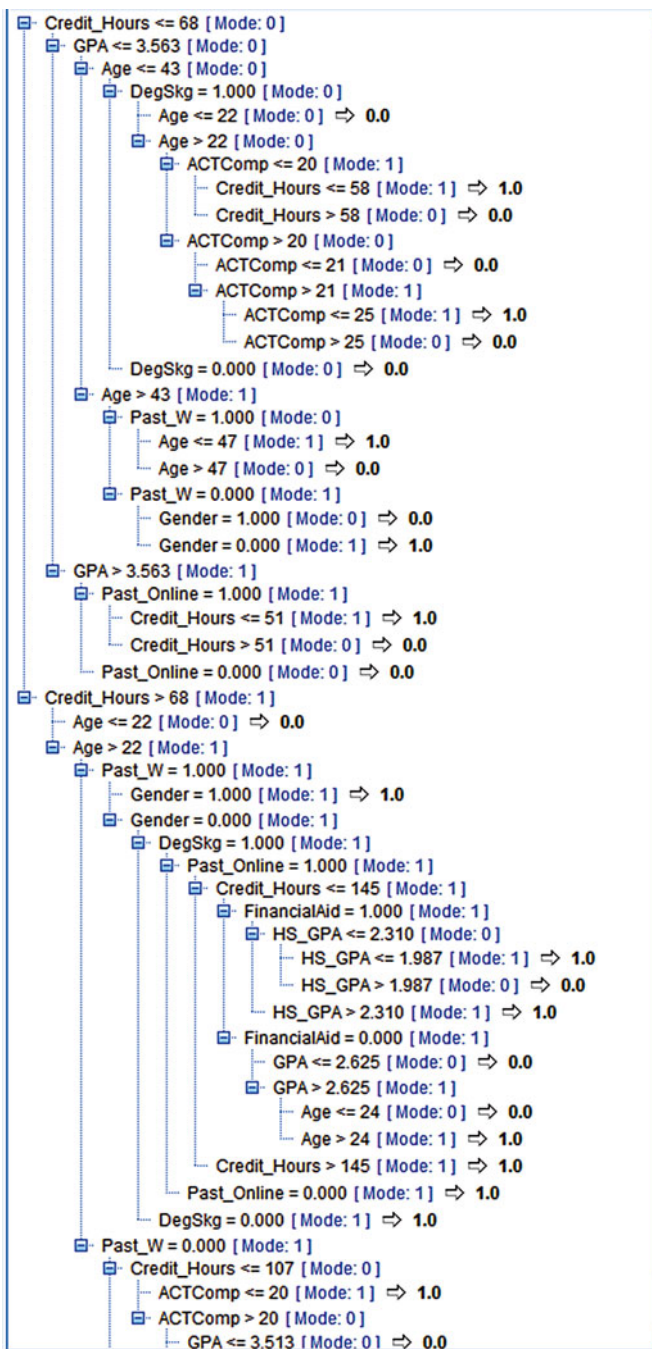
Fig. 6.2 Boosted C5.0 decision tree model for balanced SIS dataset
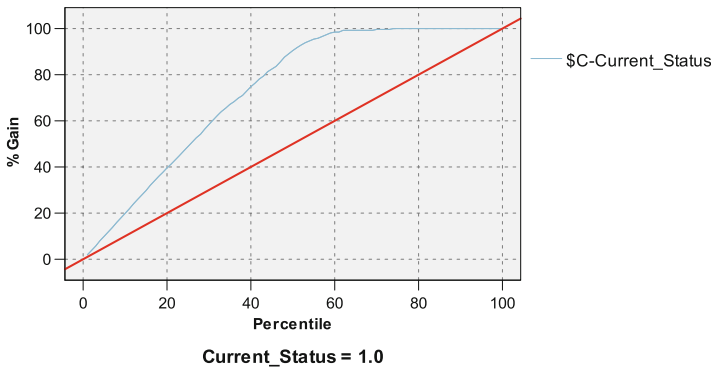
Current_Status = 1.0

Fig. 6.3 Gain chart for boosted C5.0 model

Table 6.6 Variables in the Course Management System (CMS) Dataset

No.	Variable	Data type	No.	Variable	Data type
1	Total logins	Numeric	5	Credit hours	Numeric
2	Course prefix	String	6	Days since last login	Numeric
3	Course level	Ordinal	7	Course discipline	String
4	Total time spent in minutes	Numeric	8	Course status (Dependent variable)	Binary nominal

6.4.2   CMS Data Analysis

CMS data was collected from online courses taught in the fall 2009 semester at a University located in Midwestern United States. The university uses Desire2Learn (D2L) as its Course Management System. 592 student records were collected from 29 randomly selected online courses at the end of the third day and seventh day of a course. Table 6.6 shows the CMS variables and their data types.

Although the CMS data were available from the start date of the course, this study used CMS data from the end of day 3 and end of day 7. Since the CMS did not provide automated reports suitable for data mining, a software program had to be written to extract data from individual courses for individual students. The rationale for using data from day 3 was that there was significant data available by that time, as many students did not log in before then. The rationale for using data from the end of day 7 was that in practice the institution has witnessed that most students who drop out do so within the first week of the course. There was an option to extract data at the end of day 10, but that data was not used since that date was considered late to try to prevent dropout, as many students drop out by day 10 – the institution's census date. However, it is important to note that once the model is built, an institution (including staff and instructors) could customize the model for use on any selected course dates. For example, this model could be customized to

Table 6.7 Descriptive statistics of the 3-Day CMS dataset numeric variables. n = 592

No.	Mean	Median	Std. Dev.	Skewness	Minimum	Maximum
Total logins	217.98	130	260.459	2.417	0	1,851
Course level	2.64	3	1.518	0.111	0	5
Total time spent in minutes	65.309	22.117	105.268	3.018	0	941.1
Credit hours	2.97	3	0.157	−6.056	2	3
Days since last login	1.45	1	0.951	0.577	0	3
Course status (dependent variable)	0.88	1	0.329	−2.297	0	1

run CMS data extraction on a daily basis. The descriptive statistics for the CMS data are shown in Table 6.7.

In terms of course level variable, the dataset had 28.2 % of students in courses coded as 1 (freshman), while only 3.9 % of students were enrolled in 0 level (remedial) courses. This distribution of students was expected because the institution offers relatively more 100-level courses online. In terms of the Days Since Last Login variable, only 10.3 % of students logged in to their courses on day 3. However, 57.1 % of students last logged in one day before the data extraction. In terms of the Credit Hours variable, the dataset had data for courses of only 2 and 3 credit hours where 97.5 % of students were enrolled in 3-credit hour online courses. In terms of the Discipline variable, the dataset consists of five disciplines namely: Business, Education, Humanities, Math, and Social Sciences. 28.4 % of students were enrolled in the Education discipline, which is consistent with the institution's online course offerings being primarily in the disciplines of education, business, and humanities. In terms of Course Status variable, only 12.3 % of students dropped out, while 87.7 % of students completed the courses.

The 3-day CMS dataset was analyzed using J.48 decision trees, naïve Bayes, logistic regression, MLP neural networks, and SVM. All techniques were analyzed using 10fold cross-validation. J.48 decision trees, logistic regression, and SVM provided the highest overall accuracy – 87.66 %. However, J.48, logistic regression, and SVM had low accuracy for identifying true negatives (students who actually dropped out). J.48 had only 50 % accuracy, and SVM had 0 % accuracy in identifying dropout students. All techniques performed poorly at identifying dropout students, which was understandable since the dataset had the records for only 12.5 % of students who dropped out. Since the boosted C5.0 decision tree model was significantly better with SIS data, it was worth investigating that technique with the CMS dataset. The CMS dataset was analyzed using boosted C5.0 decision trees and 10fold cross-validation. As the coincidence matrix shows (Table 6.8), the boosted C5.0 decision tree model provided an overall accuracy of 89.7 %. Additionally, it was able to identify 92.85 % of true negatives (students who actually dropped their courses). Therefore, the boosted C5.0 decision tree model was chosen for CMS data analysis due its greater overall accuracy and greater accuracy for identifying true negatives.

Table 6.8 Coincidence matrix for CMS 3-day and 7-day dataset using C5.0 boosted trees

		Dropped (0)	Completed (1)	Sum	Per class accuracy (%)	Overall accuracy (%)	ROC area
DT (C5.0 Boosted) 3-day	**Dropped (0)**	13	1	14	92.85	89.7	0.765
	Completed (1)	60	518	578	89.61		
DT (C5.0 Boosted) 7-day	**Dropped (0)**	18	5	23	78.26	89.86	0.765
	Completed (1)	5500.00 %	514	569	90.33		

The boosted C5.0 decision tree model identified total logins as the only significant predictor. This was understandable, as the other numeric variables such as days since last login and total time spent would become more significant at a later time in the course. In the boosted C5.0 decision tree, the first node was split by total number of logins. The model indicated that if the total number of logins was greater than 9, then only 9.5 % of students dropped out. If the total number of logins was less than or equal to 9 by the end of day 3, then 40 % of those students dropped out. Of those students who had less than or equal to 9 total logins by the end of day 3: 63 % of those students dropped out if the course was in the business discipline; 100 % of students in the social sciences discipline dropped out, and 23 % of students in the education, humanities, and math disciplines dropped out. Out of the 63 % of students in business courses who had total logins equal to or less than 9, 83 % dropped out if they did not log in by day 3.

In addition to analyzing the CMS data at day 3, CMS dataset from the end of day 7 was analyzed using boosted C5.0 decision trees and 10fold cross-validation. The following coincidence matrix (Table 6.8) shows an overall accuracy of 89.86 %. This model was able to accurately identify 78.26 % of true negatives, which was significantly better than boosted C5.0 with the day 3 dataset.

The boosted C5.0 decision tree model used with the 7-day CMS dataset showed more contrast in relative predictor importance. The model showed that total logins and days since last login, as well as total time spent, were strong predictors. This finding was more in alignment with what is generally witnessed in practice – that students who accessed the course less frequently, spent less total time, and had more days lapse since the last login are more prone to drop out. Course level and discipline are weak predictors of course completion or drop out.

The boosted C5.0 decision tree for the day 7 dataset had fewer decision nodes. The tree showed the first node divided between total number of logins as less than or equal to 13, or total logins as greater than 13. If total logins were greater than 13 by the end of day 7, then 90 % of those students completed their course (only about 10 % dropped out). If total logins were less than or equal to 13 by the end of day 7, then only 57 % of those students completed their course (43 % dropped out). Of the students who had 13 or less logins by the end of day 7, 78 % dropped out in the disciplines of business or social sciences, while only 25 % of those students enrolled in the humanities, math, or education disciplines dropped out.

Raw propensity scores and adjusted propensity scores, along with the predicted value of the dependent variable (Course_Status), are computed using the boosted C5.0 decision tree model. The adjusted propensity score is used to create the CMS risk score from the day 3 and day 7 datasets. Since the adjusted propensity scores are given in decimal points, they are multiplied by 100 to make them more readable. The scores are subtracted by 100 to compute the risk score for dropout. The following equation is used to compute the dropout risk score for individual students.

```
  Dynamic  CMS  Risk  Score  =IF(Total_Logins=0,"100",(100-($CRP-
Course_Status*100)))
OR
  Dynamic CMS Risk Score =IF(Total_Logins =0,"100",(100-(Adjusted
Propensity Score for Course_Status*100)))
```

6.4.3   Computation of Dynamic Risk Scores

As previously mentioned, SIS data analysis provides a baseline risk score, which can be used at the beginning of the course. As the course progresses and CMS data are analyzed, the CMS risk scores can be computed. To compute a dynamic risk score, both the baseline risk score and CMS risk scores are averaged together. The 3-day dynamic risk score is computed by averaging the baseline risk score and day 3 CMS risk score. Similarly, the 7-day dynamic risk score is computed by averaging the baseline risk score and day 7 CMS risk score.

```
  3_Day_Dynamic_Risk_Score = Baseline_ Risk_Score + 3_Day_Risk_
Score / 2
  7_Day_Dynamic_Risk_Score = Baseline_ Risk_Score + 7_Day_Risk_
Score / 2
```

For example if a student's baseline score was 20 and the 3-day CMS score was 60, then the 3-day dynamic risk score would be 40.

6.4.4   Evaluation of the Risk Scoring Model

Although the boosted C5.0 decision tree model selected for risk scoring used 10fold cross-validation, it was determined to evaluate the model with a new dataset to verify its performance and to address any possible bias. A new dataset of 205 students was extracted from the SIS and CMS. The SIS data was analyzed using a boosted C5.0 decision tree to calculate the baseline risk score from adjusted propensity scores. Similarly a CMS dataset for day 3 and day 7 was created for the 205 selected students. The CMS datasets for day 3 and day 7 were analyzed using boosted C5.0 decision trees to compute the CMS risk score for day 3 and day 7. The baseline SIS risk score and 3-day CMS risk score were then used to create the

3-day dynamic risk score; the baseline SIS risk score and 7-day CMS risk score were used to create the 7-day dynamic risk score. The data analysis for both SIS and CMS used 10fold cross-validation.

Course_Status is the actual outcome of the course, with 0 representing course dropout and 1 representing course completion. Baseline_Risk_Score is the risk score using the SIS dataset. A student is labeled at higher risk for dropout if the baseline score is 50 or higher. A baseline risk score of less than 50 suggests a lower risk of dropout (higher likelihood of course completion). All 205 records were individually reviewed to examine the baseline risk score against course status. Accuracy_Baseline_Score provides information about whether the prediction was correct or not using the following rule:

```
=IF(OR(AND(Baseline_Risk_Score>=50,Course_
Status=0),AND(Baseline_Risk_Score<50,Course_Status=1)),"Correct",
"Incorrect")
```

The above rule explains that the variable Accuracy_Baseline_Score was coded as "Correct" if Baseline_Risk_Score was greater than or equal to 50 and Course_Status was equal to 0 (dropout). Accuracy_Baseline_Score was also coded as "Correct" if Baseline_Risk_Score was less than 50 and Course_Status was equal to 1. The Accuracy_Baseline_Score variable was coded as "Incorrect" if those conditions were not met. Using the above rule, the dataset was recoded for Accuracy_Baseline_Score.

The dataset was reviewed for accuracy of the 3-day and 7-day dynamic risk scores. The variable Accuracy_3_Day_DRS (DRS stands for dynamic risk score) was used to check for the accuracy of the 3_Day_Dynamic_Risk_Score. The variable Accuracy_7_Day_DRS was used to check for the accuracy of the 7_Day_Dynamic_Risk_Score.

Accuracy_3_Day_DRS provides information about whether the prediction was correct or not using the following rule:

```
=IF(OR(AND(3_Day_Dynamic_Risk_Score>=50,Course_
Status=0),AND(3_Day_Dynamic_Risk_Score <50,Course_Status=1)),"Cor
rect","Incorrect")
```

Accuracy_7_Day_DRS provides information about whether the prediction was correct or not using the following rule:

```
=IF(OR(AND(7_Day_Dynamic_Risk_Score>=50,Course_
Status=0),AND(7_Day_Dynamic_Risk_Score <50,Course_Status=1)),"Cor
rect","Incorrect")
```

The fields of Accuracy_3_Day_DRS and Accuracy_7_Day_DRS were coded as "Correct" if the above rule conditions were met. If the rule conditions were not met, then the fields were coded as "Incorrect."

Table 6.9 provides the accuracy of the baseline risk score, the 3-day dynamic risk score and the 7-day dynamic risk score in the evaluation dataset. As indicated in the table, 82.4 % of records were correctly predicted using the baseline risk score whereas the accuracy of the baseline and 3-day dynamic score was 78.5 %, which was less than the accuracy of the baseline risk score. This lower accuracy of the day 3 dynamic risk score was a result of lower performance of the 3-day CMS dataset – indicating that the extraction of the end of day 3 CMS data does not improve the

Table 6.9 Accuracy of baseline, 3-day dynamic, and 7-day dynamic risk scores

	Baseline risk score		3-day dynamic risk score		7-day dynamic risk score	
	Using SIS evaluation dataset		Using SIS and 3-day CMS Evaluation datasets		Using SIS and 7-day CMS Evaluation datasets	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Frequency	169	36	161	44	177	28
Percent	82.4	17.6	78.5	21.5	86.3	13.7

Table 6.10 Comparison of pre-census and post-census dropout accuracy

Dropout accuracy	Baseline risk score (score 50 or above)	3-day dynamic risk score (score 50 or above)	7-day dynamic risk score (score 50 or above)
Pre-census	91.3 % (42/46)	43.47 % (20/46)	63.04 % (29/46)
Post-census	76.92 % (20/26)	57.69 % (15/26)	88.46 % (23/26)

baseline accuracy provided by the SIS dataset. The table also includes the accuracy of the 7-day dynamic risk score, which is derived from the baseline risk score and the 7-day CMS risk score. The results show that the accuracy of the 7-day dynamic risk score was 86.3 %, which was higher than the baseline risk score and 3-day dynamic risk score. This indicates that by day 7 the CMS has sufficient data to improve on the accuracy of the overall prediction.

6.4.5 Comparison of Pre-census and Post-census Dropout Accuracy

Since the university provides a full refund to students who drop out before the census date, it would be important to investigate how accurate risk scores are for students who drop out before and after the census date. There were 72 records for dropout students (out of a total of 205 records) in the evaluation dataset. 46 of those 72 students dropped before the census date, while the remaining 26 dropped after the census date.

Table 6.10 explains the before census dropout accuracy and after census dropout accuracy for the baseline risk score, 3-day dynamic risk score, and 7-day dynamic risk score. The baseline risk score was found to be 91.3 % accurate in identifying students who dropped out before the census date. The accuracy of the 7-day dynamic risk score was highest at 88.46 % for students who dropped out after the census date. The 3-day dynamic risk score was marginally accurate for both pre- and post-census dropout – which is likely to be related to insufficient student activity through day 3. Based on this analysis, it is evident that both the baseline risk score and 7-day risk score are important for the accurate identification of dropout students.

6.4.6 Recommendation for Deployment of Predictive Models

Deployment is the final phase in the CRISP-DM process. The data analysis and evaluation phases indicated that the boosted C5.0 decision tree model was most suitable for both SIS and CMS datasets to predict at-risk students and to compute baseline and dynamic risk scores. To deploy the boosted C5.0 SIS and CMS predictive models, IBM Modeler and a database server, such as SQL server, can be used. IBM Modeler has the capability to export the model in PMML, which can be loaded to a database server to score new datasets. A web-based interface connected with a back-end data source containing risk scores can serve as the primary user interface to find risk scores for students. The risk scores and decision rules from boosted C5.0 decision trees can serve as rules for triggering alerts. For example, a rule could be set with the provision that if the predictive model has a risk score for a student of over 80, then an e-mail alert could be triggered to the instructor and retention personnel.

The purpose of early alerts and meaningful recommendations that help provide early intervention is to prevent dropout before it occurs. This approach has been noted in key retention studies (Seidman 2005). This study seeks to build an early alert system and recommender system, as a possible application of the predictive models, which use background research in recommender systems and student retention. The Seidman Retention formula based on the Tinto's model (Tinto 1987, 1993). Tinto's model discusses the role of early identification and intervention in improving student retention. The formula states:

```
Retention = early identification + (early + intensive + continuous)
intervention
```

The early identification of at-risk students allows for early intervention. The predictive models and risk scores developed in the previous section help identify students who are at risk of dropping out. They can provide insights that can be developed into recommendations for intervention. Accurate early identification of at-risk students, with effective corresponding recommendations, is helpful for intervention that is early, intensive, and continuous. Early identification of at-risk students is achieved through the SIS based predictive model (predictive model I) and baseline risk score when the student enrolls in a course. As the course progresses and the CMS predictive model (predictive model II) and dynamic risk score become available, there is an opportunity for the continuous monitoring of at-risk students. Predictive model I and predictive model II allow early identification of at-risk students and can help create student-specific recommendations, leading to early, continuous, and effective interventions. In this systemic view:

```
Intervention (early + intensive + continuous) = At-risk student
identification (early + accurate + continuous) + student-specific
recommendation
```

Figure 6.4 shows the predictive analytics based intervention.

This study presents a framework for a recommender system, which uses rule-based predictive analytics. There are additional steps in deployment for creating early alerts and recommendations using the predictive models discussed in this section. The deployment steps for the recommender system are discussed in the next section.

Fig. 6.4 Predictive analytics
based intervention
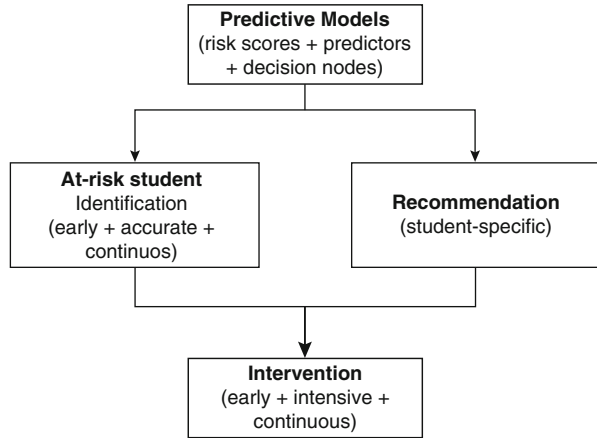
6.4.7 Recommender System Architecture

The recommender system constitutes student rules, system rules, and a recommendation repository. Student rules and system rules are used for the early alert system that can be combined with the appropriate recommendation from the recommendation repository. A system rule, for example, is the minimum risk score (determined by a preselected value by the institution) at which an alert will be triggered. The alert system will send e-mails to retention personnel, the instructor, and the student that will include recommendations from the recommendation repository. If the student has not previously taken an online course, that student will be recommended web-based CMS tutorials and links to student support services in the e-mail. If a student does not have financial aid, the early alert system can send an e-mail alert to financial aid and retention personnel for the purpose of apprising the student of relevant information and/or assistance. Previous alerts and recommendations can be stored in a repository so they can be tracked by instructors and retention personnel using a web-based interface. The recommender system rules, the predictive model, and the risk score/recommendation repository are used to build the web-based interface of the recommender system. Such a system can be built using a back-end database server, such as SQL server or Oracle, with a web programming language (PHP, ASP.net or Java Server Pages, etc.). Figure 6.5 presents the general recommender system architecture.

6.5 Conclusion

This study identified four primary research objectives. Each objective of the study was met, as described below.

1. Identify and analyze various SIS and CMS based predictors relative to their strengths to predict dropout risk for students in online courses.
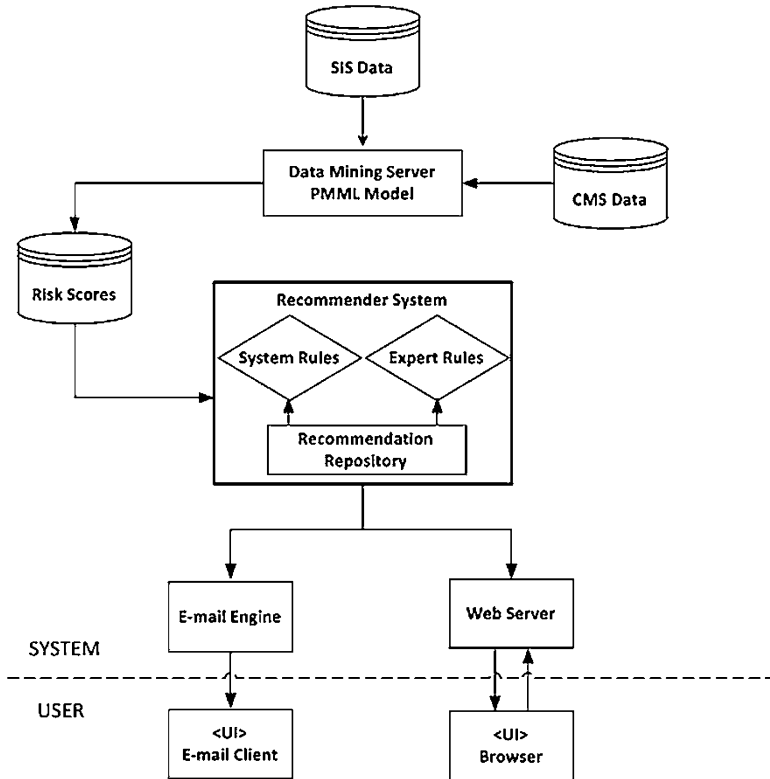
Fig. 6.5 Recommender system architecture

This study identified 10 SIS independent variables using the literature review that have been shown to have an effect on student dropout. This study concludes that all SIS variables play a role in student dropout. Furthermore, the study identified 7 CMS variables including: total logins, course prefix, course level, total time spent, course credit hours, days since last login, and course discipline. The analysis showed that the end of day 7 CMS dataset can provide sufficient information to develop a predictive model. The analysis showed that total logins, days since last login, and total time spent were significant predictors.

2. Evaluate various statistical and machine learning techniques (such as Neural Networks, SVM, and hybrid methods) for their predictive accuracy to build predictive models to leverage both SIS (historical) and CMS (time-variant) data.

Based on this comparative analysis, it was concluded that the C5.0 decision tree algorithm performed the best. The predictive performance of this classifier can be further enhanced using boosting. The boosted C5.0 decision tree models provided the best accuracy and performance for both the SIS and CMS datasets.

3. Construct a baseline predictive model using SIS variables and dynamic predictive model using both SIS and CMS variables to model dynamic dropout risk in the form of a risk score.
4. Propose a recommendation system architecture based on the predictive models to identify at-risk students and to offer meaningful alerts and recommendations based on their risk score.

Biography

Rajeev Bukralia is Associate Lecturer with the Information and Computing Science Department at the University of Wisconsin-Green Bay. His previous positions include serving as the Associate Provost for Information Services and CIO at the University of Wisconsin-Green Bay and Dean of Educational Outreach and Libraries at Black Hills State University. He holds a Doctor of Science degree in Information Systems from Dakota State University. His research interests include predictive and prescriptive analytics, machine learning, data science, and e-learning.

Amit V. Deokar is an Assistant Professor of Information Systems in the Sam and Irene Black School of Business at Pennsylvania State University, Erie, Pennsylvania. His recent research interests are in decision support and analytics, business process management, and collaboration processes and technologies. He has published several conference publications, journal articles, and book chapters in these areas. He holds a BE in Mechanical Engineering from V.J. Technological Institute, Mumbai, a MS in Industrial Engineering from the University of Arizona, and a PhD in Management Information Systems from the University of Arizona. He is a member of AIS, INFORMS, ACM, and AAAI.

Surendra Sarnikar is an Associate Professor in Information Systems at the College of Business and Information Systems, Dakota State University. He holds a Bachelors degree in Engineering from Osmania University, India, and a PhD in Management Information Systems from the University of Arizona. He teaches healthcare informatics, design research and knowledge management at the Dakota State University. He has published several conference and Journal publications in the area of healthcare information systems, knowledge management systems, and information retrieval.

References

Allen, J., & Robbins, S. B. (2008). Prediction of college major persistence based on vocational interests, academic preparation, and first-year academic performance. *Research in Higher Education, 49*(1), 62–79.

Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computers & Education*, *53*(3), 563–574. Retrieved from http://www.sciencedirect.com/science/article/B6VCJ-4W4S30Y-2/2/fbc3488dc25814084c5e415f59c859b9

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485–540.

Bellaachia, A., Vommina, E., & Berrada, B. (2006). *Minel: A framework for mining e-learning logs* (p. 263). Anaheim: ACTA Press.

Braxton, J. M. (2000). *Reworking the student departure puzzle*. Nashville: Vanderbilt University Press.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1). Belmont: Wadsworth.

Bukralia, R., Sarnikar, S., & Deokar, A. V. (2009). Predictive modeling to improve retention of online students. In *Proceedings of the 4th Midwest association for information systems conference (MWAIS'09)*, Madison.

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education, 64*(2), 123–139.

Campbell, J. (2008). *Analysis of institutional data in predicting student retention utilizing knowledge discovery and statistical techniques*. Arizona: Northern Arizona University.

Campbell, P. J., & Oblinger, D. (2007). *Academic analytics*. EDUCAUSE. Retrieved from http://net.educause.edu/ir/library/pdf/PUB6101.pdf

Campbell, J. P., Finnegan, C., & Collins, B. (2006). Academic analytics: Using the CMS as an early warning system. *WebCT impact conference 2006*, Chicago.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a New Era. *Educause Review*, *11*, 41–57.

Carr, S. (2000). As distance education comes of age, the challenge is keeping the students: Colleges are using online courses but retaining them is another matter. *The Chronicle of Higher Education*, 41–57. Retrieved from http://chronicle.com/free/v46/i23/23a00101.htm

Chen, C. M., Chen, Y. Y., & Liu, C. Y. (2007). Learning performance assessment approach using web-based learning portfolios for e-learning systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37*(6), 1349–1359.

Cocea, M., & Weibelzahl, S. (2007). Cross-system validation of engagement prediction from log files. In *Creating new learning experiences on a global scale* (pp. 14–25). Berlin: Springer.

Dawson, S., McWilliam, E., & Tan, J. P. L. (2008). Teaching smarter: How mining ICT data can inform and improve learning and teaching practice. *Hello! Where are you in the landscape of educational technology? Proceedings ascilite Melbourne 2008*, Deakin University, Melbourne.

Denson, K., & Schumacker, R. E. (1996). Student choices: Using a competing risks model of survival analysis. Annual Meeting of the American Educational Research Association. New York, NY

Diaz, D. (2000). *Comparison of student characteristics, and evaluation of student success, in an online health education course.* Nova Southeastern University. Retrieved from http://www.ltseries.com/LTS/pdf_docs/dissertn.pdf

Dutton, M., & Perry, J. J. (2002). Do online students perform as well as lecture students? *Journal of Engineering Education, 90*(1), 131.

Etchells, T. A., Nebot, A., Vellido, A., Lisboa, P. J. G., & Mugica, F. (2006). Learning what is important: Feature selection and rule extraction in a virtual course. In *The fourteenth European symposium on artificial neural networks* (pp. 401–406). Bruges: Citeseer.

Hegedorn, L. (2005). How to define retention. In A. Seidman (Ed.), *College student retention: Formula for student success*. Westport: Greenwood Publishing.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression. *New Directions for Institutional Research, 133,* 17–33.

Hung, J.-L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching, 4*(4), 426–437.

Kemp, W. C. (2002). Persistence of adult learners in distance education. *American Journal of Distance Education, 16*(2), 65–81.

Kiser, A., & Price, L. (2007). The persistence of college students from their freshman to sophomore year. *Journal of College Student Retention: Research, Theory and Practice, 9*(4), 421–436.

Lotkowski, V. A., Robbing, S. B., & Noeth, R. J. (2004). *The role of academic and non-academic factors in improving college retention*. Iowa City: American College Testing. Retrieved from http://inpathways.net/college_retention.pdf

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53*(3), 950–965. Retrieved from http://www.sciencedirect.com/science/article/B6VCJ-4WDNBR1-3/2/8f115f84202d1a7ee5ae97155efb4b44

Lynch, M. M. (2001). Effective student preparation for online learning. The Technology Source, November/December 2001. Retrieved from http://www.technologysource.org/article/100/

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599. Retrieved from http://www.sciencedirect.com/science/article/B6VCJ-4XBF8RB-4/2/7fc8b75914be9b80d943da1d37793f06

Mangold, W. D., Bean, L. A. G., Adams, D. J., Schwab, W. A., & Lynch, S. M. (2002). Who goes who stays: An assessment of the effect of a freshman mentoring and unit registration program on college persistence. *Journal of College Student Retention: Research, Theory and Practice, 4*(2), 95–122.

Morris, L. V, Finnegan, C., & Wu, S.-S. (2005). Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, *8*(3), 221–231. Retrieved from http://www.sciencedirect.com/science/article/B6W4X-4GX6J8V-4/2/be57389e28766a8f9d0257a011719be1

Muehlenbrock, M. (2005). Automatic action analysis in an interactive learning environment. In *Twelfth international conference on artificial intelligence in education*, Amsterdam.

Muse, H. E. (2003). The web-based community college student: An examination of factors that lead to success and risk. *The Internet and Higher Education, 6*(3), 241–261.

Newell, C. (2007). *Learner characteristics as predictors of online course completion among nontraditional technical college students*. University of Georgia. Retrieved from http://www.coe.uga.edu/leap/adminpolicy/dissertations_pdf/2007/newell_2007_edd.pdf

Nisbet, R., Elder, J. F., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Burlington: Academic.

O'Brien, C., & Shedd, J. (2001). *Getting through college: Voices of low income and minority students in New England*. Washington, DC: The Institute for Higher Education Policy.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*(1), 169–198.

Osborn, V. (2001). Identifying at-risk students in videoconferencing and web-based distance education. *American Journal of Distance Education, 15*(1), 41–54.

Park, J. H. (2007). Factors related to learner dropout in online learning. In *International research conference in The Americas of the academy of human resource development*. Indianapolis. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED504556

Park, J. H., & Choi, H. J. (2009). Factors influencing adult learners' decision to drop out or persist in online learning. *Educational Technology & Society, 12*(4), 207–217.

Parker, S., & Greenlee, H. (1997). From numbers to action: a preliminary study of retention. In *Annual Forum of the Association for Institutional Research*. Albuquerque.

Pittman, K. (2008). *Comparison of data mining techniques used to predict student retention*. Florida: Nova Southeastern University.

Porter, O. F. (1989). *Undergraduate completion and persistence at four-year colleges and Universities: Completers, Persisters, Stopouts, and Dropouts*. Washington, DC: National Institute of Independent Colleges and Universities.

Roblyer, M. D., Davis, L., Mills, S. C., Marshall, J., & Pape, L. (2008). Toward practical procedures for predicting and promoting success in virtual school students. *American Journal of Distance Education, 22*(2), 90–109.

Ross, L. R., & Powell, R. (1990). Relationships between gender and success in distance education courses: A preliminary investigation. *Research in Distance Education, 2*(2), 10–11.

Rovai, A. P. (2003). In search of higher persistence rates in distance education online programs. *The Internet and Higher Education*, *6*(1), 1–16. Retrieved from http://www.sciencedirect.com/science/article/B6W4X-47HC8MR-1/2/f098c4771634ad39b2de7225c40e5e76

Seidman, A. (2005). Where we go from here: A retention formula for student success. In A. Seidman (Ed.), *College student retention* (p. 296). Westport: Praeger Publishers.

Tinto, V. (1987). *The principles of effective retention*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED301267

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition*. Chicago: University of Chicago Press.

Vare, J. W., Dewalt, M. W., & Dockery, R. E. (2000). Predicting student retention in teacher education programs. In *Proceedings of the annual meeting of the American association of colleges for teacher education*. Chicago.

Wang, A. Y., & Newlin, M. H. (2002). Predictors of performance in the virtual classroom: Identifying and helping at-risk cyber-students. *The Journal of Higher Education Academic Matters, 29*(10), 21–25.

Whiteman, J. M. (2004). *Factors associated with retention rates in career and technical education teacher preparation web-based courses*. Orlando: University of Central Florida.

Willging, P., & Johnson, S. (2004). Factors that influence students' decision to dropout of online courses. *Journal of Asynchronous Learning Networks, 8*(4), 105–18.

Yu, C. H., DiGangi, S. A., Jannasch-Pennell, A., Lo, W., & Kaprolet, C. (2007). A data-mining approach to differentiate predictors of retention. In *EDUCAUSE Southwest conference*. Austin. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED496657

Chapter 7
Membership Reconfiguration in Knowledge Sharing Network: A Simulation Study

Suchul Lee, Yong Seog Kim, and Euiho Suh

Abstract The purpose of this study is to propose a new approach that minimizes the negative impacts of structural barriers to knowledge sharing in the current of knowledge sharing networks by dynamically reconfiguring communities of practice (CoP) memberships. For this purpose, we develop several propositions to determine source CoPs, destination CoPs, rearrangement candidates, and recipient candidates to regulate the process of reconfiguring collaboration networks of source CoPs and reconstructing networks of destination CoPs after reallocating members from source CoPs to destination CoPs. To test the validity and usefulness of the proposed approach, we simulate two reconfiguration strategies that are different in the sense whether or not the distribution of expertise levels of CoP members is considered to determine the destination CoP. Our experimental results confirm that the proposed approach with either strategy effectively decreases potential threats to collaboration among CoP members and improves the structural healthiness of knowledge sharing networks of departments and organization. In particular, the number of CoPs in which knowledge creating is more active than knowledge sharing is significantly increased while the number of inactive CoPs is decreased. We attribute this finding to the fact that both experts and non-experts members are more evenly distributed across CoPs through rearrangement and these experts with light collaboration burden post their knowledge and practical skills to help non-experts in their CoPs.

Keywords Knowledge sharing network • Communities of practice • Knowledge management system • Membership reconfiguration • Bottleneck impact score (BIS)

S. Lee
Department of Future R&D Strategy, Division of Policy Research,
Korea Institute of Science and Technology Information (KISTI),
245 Daehang-ro, Yuseong-gu, Daejeon 305-806, Republic of Korea
e-mail: quito@postech.ac.kr

Y.S. Kim (✉)
MIS Department, Jon M. Huntsman School of Business, Utah State University,
Logan, UT 84322-3515, USA
e-mail: yong.kim@usu.edu

E. Suh
Department of Industrial and Management Engineering, Pohang University of Science
& Technology, 77 Cheongam-Ro, Nam-Gu, Pohang, Gyeongbuk, Republic of Korea
e-mail: suchul.lee@kisti.re.kr

7.1 Introduction

Recently, many organizations strive to integrate and maximize the use of knowledge and best practices embodied in the expertise of experts by operating knowledge management systems (KMS) (Fritz et al. 1998; Griffith et al. 2003). According to many studies, KMS not only enhances knowledge reuse but also stimulates innovative solutions by turning members' intellectual capital into knowledge resources that will improve the organization's capacity to cope with increased levels of competition and shortages of qualified knowledge workers (Janz and Prasarnphanich 2003; Von Krogh 1998). In particular, many practitioners and scholars are paying increasing attention to communities of practice (CoPs), an informal and spontaneous network of organizational members toward the common goal of sharing knowledge and best practices to solve problems (Brown and Duguid 1991; McDermott 1999). It is believed that CoPs help organizations not only inspire members to use their talents and best practices but also facilitate and revise new organizational strategies by allowing members to constantly exchange, validate, and refine multiple perspectives on work-related problems and issues (Lesser and Storck 2001; McDermott and Archibald 2010).

While CoPs are self-emerging and self-organizing networks in their nature, they are unlikely to be successful unless organizations cultivate environments in which members are strongly encouraged to share their knowledge by eliminating any structural bottlenecks or psychological barriers (Helms 2007; Helms et al. 2010; Lee et al. 2012). For this reason, it is not difficult to find formally and informally formed CoPs, and more organizations are interested in assessing the structural health of their CoPs to remove bottlenecks to employees' knowledge-sharing activities. One of the most well-known and successful treatments is to motivate organizational members by providing intrinsic and extrinsic (e.g., financial) rewards for actively engaging CoP members and CoPs (McDermott 1999). In this study, we like to boost knowledge sharing activities among CoP members by dynamically reconfiguring CoP memberships (i.e., reallocating members from a CoP (source CoP) to another CoP (destination CoP)) to minimize the negative impacts of loosely connected structure of CoP networks and any existing bottlenecks in the current CoP networks. Ideally these two methods—an organizational human resource management approach and a social network structural approach—can be combined to obtain optimal results.

As a prerequisite of our approach, the management teams should quantitatively diagnose whether any threats to knowledge management initiatives exist in their organizations and how serious they are. To this end, we rely on a bottleneck impact score (BIS) metric (Lee et al. 2012) that is a weighted sum of the pervasiveness of six bottlenecks in two possible barrier categories: master-apprenticeship relations and knowledge drain. Then we develop several propositions to determine ideal source and destination CoPs for reconfiguration and ideal candidates for reallocation in source CoPs. In addition, several other propositions are devised to regulate the process of reconstructing collaboration networks among remaining members in source CoPs and between new members and old members in destination CoPs.

We validate the effectiveness of our reconfiguration strategies by measuring the improvements of total sum of BISs. In terms of methodology, we combine a simulation approach and a social network analysis (SNA) based on real-world CoP datasets. Since the proposed approach continuously reconfigures the structure of knowledge sharing networks by dynamically reconfiguring CoP memberships to reduce master-apprenticeship relations and knowledge drain barriers, the resulting structure should have a decreased value of BIS metric and minimize the losses due to the business discontinuities caused by such risks.

This paper is organized as follows. In the following sections, we first briefly review several relevant studies that provide theoretical and empirical grounds for this study. Then we describe the framework of our KMS with important propositions to reconfigure collaboration networks among CoP members. Data pre-processing and simulation experimental setups are immediately followed. Experimental outcomes are presented and discussed in terms of BIS metric and improvement of structural changes in collaboration networks. Finally, we provide concluding remarks and suggestions for future research.

7.2 Literature Review

7.2.1 Social Behavior Theories, Collaboration Climates, and Knowledge Sharing

Several social behavior theories–social exchange theory (Baum et al. 2001; Blau 1964; Molm 1997), expectancy theory (Vroom 1964; Wang and Strong 1996), public goods theory (Fulk et al. 2004; Marwell and Oliver 1993), and social capital theory (Deci 1971; Nahapiet and Ghoshal 1998; Nebus 2004; Putnam 1995)–are often adopted to understand and explain knowledge sharing behaviors of organizational members. Among those, public goods theory and social capital theory explain organizational members' conflicting perspectives on knowledge sharing activities under knowledge management including KMS and CoPs. First, public goods theory regards knowledge in pubic place such as KMS as one of public goods and raises the free-rider problem in that individuals who do not contributed to the creation of knowledge bases can benefit from accessing KMS (Fulk et al. 2004; Marwell and Oliver 1993) and hence some individuals may withhold tacit knowledge for themselves only (Bock et al. 2006; Deci 1971; Thorn and Connolly 1987; Venkatesh and Davis 2000). However, according to social capital theory, individuals may not want to be a free rider but, instead, they like to invest in social relations, the resources tied up in those connections, and the ability of securing benefits from those relationships (Borgatti and Foster 2003; Kilduff and Tsai 2003). Therefore, individuals want to build high social capital by not only reusing knowledge from the KMS but also contributing knowledge to it over time (Bock et al. 2008). Individuals with high social capital are willing to share various information across groups, engage in problem solving, and actively collaborate with others to get work

done (Cross and Parker 2004; Dalkir 2005). Because of those organizational members' conflicting perspectives on knowledge sharing activities, many organizations has been faced with some barriers to reach success of knowledge management, and thus they have focused on cultivating organizational knowledge sharing climate which make organizational members follow social capital theory instead of public good theory.

Several studies identify collaboration climate in organizations as a critical success factor of knowledge sharing (Constant et al. 1996; Huber 2001; Orlikowski 1993). Scholars in cross-cultural research also argue that cultural factors such as group conformity and face saving in a Confucian society can directly affect intention to collaboration (Bang et al. 2000; Tuten and Urban 1999). For example, fair organizational practices build trust between members and lead employees to go beyond the call of duty to share their knowledge (Kim and Mauborgne 1997). Similarly, Al-Alawi et al. (2007) identify trust, communication, and rewards as critical organizational culture for successful knowledge sharing, and suggest to cultivate appropriate climates by arranging social events and outdoor discussions, providing sufficient information systems, and providing effective rewards. Individuals in innovative and pro-social work context are more likely to share new and creative ideas with each other and encourage a sense of collaboration among members (Kim and Lee 1995). Another research (Bock et al. 2005) recognizes a climate of trust, tolerance of failure, and pro-social norms as three organizational factors for successful knowledge sharing. According to Gupta (2008), employees with lower job-levels show higher integrity, respect, and trust than employees with higher job-levels such as executives, encouraging the need to cultivate knowledge sharing climate for employees with higher job-levels.

Interestingly, Hinds and Kiesler (1995) argue that technical workers (e.g., software engineers) rely extensively on lateral communication in CoPs because of the nature of the work they perform and the way they are organized. Similarly, Ahuja and Carley (1999) posit that non-routine tasks can be better performed through lateral communication and under the nonhierarchical coordination form, resulting in strongly tied network structure with active knowledge sharing activity among members. It is also shown that external factors such as institutional structures influence the salience of subjective norms (Bearden and Etzel 1982; Lee and Green 1991; Triandis 1972; Tse et al. 1988). For example, organizational incentive structures such as pay-for-performance compensation schemes can discourage knowledge sharing if employees believe that knowledge sharing will hinder their personal efforts to distinguish themselves relative to their coworkers (Huber 2001). Finally, collaborative climate seems to better in the small- to mid-size organization than in large organization, and in the private sector than in the public sector (Sveiby and Simons 2002).

7.2.2   Knowledge Sharing Bottlenecks and BIS Metric

As the importance of CoP has been emphasized, some researchers have suggested a diagnosis or an evaluation methodology for CoP activities (Botkin 1999; Lesser and Storck 2001; McDermott 1999; Wenger and Snyder 2000; Zhang and Watts 2008).

Since the first introduction of social network graph (Moreno 1934), the sociogram that contains actors as nodes and their relationships as links between the nodes, has been used to understand knowledge sharing activities. For example, Cross et al. (2000) employed SNA to visualize and understand the multitude of social relationships among members that can either facilitate or impede knowledge sharing. Specifically, they analyzed members' understanding of each other's knowledge to assess the overall cohesion of the group, and identified the core members and isolated members in the network like other studies (Cantner and Graf 2006; De Laat et al. 2007; Haythornthwaite 1996). In Bosua and Scheepers (2007), the Bosua-Scheepers Model (BSM) was introduced for an assessment of knowledge sharing activity assuming that efficient and effective knowledge sharing occurs only if current networks have an appropriate maturity and are supported by facilitating mechanisms such as email and online meetings. Finally, a recent study (Iyengar et al. 2011) showed that when low-status individuals are clustered around high-status individuals, they are more likely to engage in social dynamics than when their cluster is distantly separated from a densely connected core of high-status individuals.

However, more relevant analyses in regard to diagnosing the structural healthiness in terms of knowledge networks in CoPs can be found in Lee et al. (2012) in which CoPs are classified into four types: knowledge sharing community (CoP^{SH}), knowledge storing community (CoP^{ST}), knowledge learning community (CoP^{LR}), and inactive community (CoP^{IA}). In short, CoPs categorized as CoP^{SH} perform active knowledge sharing activities in both creating and consuming, and CoP^{IA} includes CoPs whose knowledge creation and consumption activities are inactive. If a CoP is classified into neither CoP^{SH} nor CoP^{IA} and more interested in knowledge creating than consuming, then it is classified into CoP^{ST} while a CoP which has opposite trend is identified as CoP^{LR}. More importantly, they also investigate whether or not there are any structural weaknesses in knowledge networks by identifying the existence and seriousness of two major barriers, master-apprenticeship and knowledge drain barriers. According to Lee et al. (2012), the master-apprenticeship barrier includes four types of bottlenecks depending on the characteristics of the links between experts and non-experts. The first bottleneck (Bottleneck 1) addresses a case in which experts engage in knowledge transfer with too few non-experts (fewer than two), while the second (Bottleneck 2) refers to a case in which non-experts learn their best practices from too few experts (fewer than two). In the last two (Bottlenecks 3 and 4), experts engage in knowledge transfer with too many non-experts (more than four), and non-experts learn best practices from too many experts (more than four experts), respectively. In contrast, knowledge drain barrier recognizes knowledge drain risk that becomes an issue when experts who maintain few or none connections leave the organization (Bottleneck 5) or when organizational members who are not necessarily experts but who maintain high connectivity with others leave the organization (Bottleneck 6). Finally, they measure the pervasiveness of such bottlenecks using a bottleneck impact score (BIS) metric defined as $BIS = \sum_i BIS_i = \sum_i w_i \cdot p_i$ where w_i and p_i represent the relative priority and pervasiveness of the ith bottleneck, respectively. The same definitions of bottlenecks and BIS are adopted for this study.

Another relevant study (Kwon et al. 2007) investigated several ontological forms of network structures and evaluated the structural efficiency and stability embedded in each identified network under organizational downsizing through computer simulation. Particularly, they explored four ontological social network archetypes–random, small world, moderate scale free (MSF), and high scale free (or Barabasi)–and found that centralized coordination structures such as MSF and Barabasi are generally more resilient and facilitate better coordination to preserve a worker's efficiency and the stability of the network structure under a relatively small-scale workforce reduction. To this end, they proposed two alternative reconnecting mechanisms in the face of downsizing: "planned" or "unplanned" tactics (Ahuja and Carley 1999). In the case of "planned" tactics, the tasks performed by departed members prior to downsizing are reassigned to the remaining members with maximum capacity because that structural change is well designed, fully planned, and smoothly executed, while organizations randomly reassign disconnected wires to existing nodes in the case of "unplanned" tactics because they are not well prepared for workforce shrinkage. While their planned tactics are adopted in our study with minor changes to reconfigure collaboration networks in source CoPs, this study takes one step further by proposing tactics to create new connections of departed members from source CoPs with members in destination CoPs.

7.3 Framework of the Proposed KMS

7.3.1 Rearrangement Propositions of CoP Members

The proposed KMS is based on the fundamental assumption that dynamically rearranging redundant CoP members with high level of knowledge from highly performing CoPs (e.g., CoP^{SH}) to poorly performing CoPs (e.g., CoP^{IA}) improve the efficiency of knowledge sharing in both CoPs by (at least partially) eliminating existing bottlenecks. Note that while CoPs are characterized as informal and self-organizing, they can be nourished by strategically *seeding* active members who are willing to share their knowledge (Wenger and Snyder 2000). Ultimately, this improvement will result in the efficiency of bilateral communications and exchanges of knowledge and experiences among organizational members.

To this end, we posit several rearrangement propositions that regulate rearrangement process of CoP members across CoPs in this study. Note that these propositions were suggested to CoP management teams in Company P as a strategic approach to enhancing knowledge sharing activities and modified to reflect feedbacks in terms of organizational and technical feasibility.

For notation convenience, we denote a CoP where an expert or specialist member is selected for rearrangement purposes and a CoP where a new member is assigned into as source and destination CoP, respectively. We also denote CoP members in the destination CoP who are going to be connected with the rearrangement candidate from the Source CoP as recipient candidates.

The first set of propositions regulates the selection of the source CoP and the rearrangement candidate member within the source CoP, and we define them as follows:

Proposition 1-a *Selection Strategy of the Source CoP*. The ideal candidate for the source CoP is one of CoPs with the highest value of BIS_1 (i.e., where experts engage in knowledge transfer with too few non-experts) or BIS_4 (i.e., where non-experts learn best practices from too many experts).

Proposition 1-b *Selection Strategy of the Rearrangement Candidate in the Source CoP*. The ideal candidate for the rearrangement candidate is an expert or a specialist member who engage in knowledge transfer with too few non-experts. However, if the candidate is the only expert or specialist in the source CoP, then the member is not selected.

The reasoning behind for Proposition 1-b is that the members who perform core activity is one of the most critical ingredient for the growth of communities (Jones et al. 2004) and they can bring a broad span of influence in CoPs (Blyler and Coff 2003). It is believed that these seeding members rearranged into poorly performing CoPs are most likely to arouse knowledge sharing activities among members and, if successful, inactive members are likely to show herding behavior by imitating what active members are doing (Oh and Jeon 2007). Ultimately, the well-distributed *seeding* active members across CoPs may act as catalysts to build a favorable organizational climate for knowledge sharing.

Note also that we only consider master-apprentice bottlenecks in the selection process of the source CoP and the rearrangement candidate mainly because knowledge drain bottlenecks cannot be directly controlled by rearrangement strategies. However, both master-apprenticeship bottlenecks and knowledge drain bottlenecks are fully considered when the outcome of rearrangement strategy is estimated in terms of *BIS* to accurately estimate the structural risk of CoP network based on all identifiable bottlenecks.

Once the source CoP and the rearrangement candidate are selected, it is necessary to determine the destination CoP and recipient candidates so that the rearrangement candidate can be reconfigured with recipient candidates in the destination CoP to maximize the impact of rearrangement strategy. To this end, another set of propositions regulates main and supplemental strategy to select the ideal destination CoP and the recipient candidate. These propositions are formally specified as follows:

Proposition 2-a *Main Selection Strategy of the Destination CoP*. The destination CoP must belong to the same department of the source CoP and should not be one of CoPs that the rearrangement candidate has a membership.

Proposition 2-b *Supplemental Selection Strategy of the Destination CoP*. Among CoPs that satisfies the requirement specified in Proposition 2-a, the ideal candidate for the destination CoP is one of CoPs with the highest value of BIS_2 (i.e., non-experts learn best practices from too few experts) or BIS_3 (i.e., where experts engage in knowledge transfer with too many non-experts).

Proposition 2-c *Selection Strategy of the Recipient Candidate in the Destination CoP*. The ideal candidate for the recipient candidate is an expert (or a specialist member) who most actively engages in knowledge transfer with (and possibly overwhelmed by) non-experts.

The reasoning behind Proposition 2-a is that when a rearrangement candidate moves to the destination CoP in a different department (e.g., Iron & Steel department to Staff department), she is most likely to remain inactive mainly because she has not accumulated knowledge and experiences relevant and useful to members in another department. In addition, reallocating a member into another CoP in the same department is likely to remove unnecessary times to adapt to other members and their communication patterns in other departments. Another special case to remind is that a rearrangement candidate engages in both the source CoP and the destination CoP because an employee can participate in multiple CoPs. In this case, CoPs that already include the candidate as a member cannot be selected as a destination CoP. To complete a rearrangement strategy of the rearrangement candidate from the source CoP to the destination CoP, it is necessary to first reconfigure network connections that the rearrangement candidate has maintained with other members in the source CoP. To this end, we present two propositions as follows:

Proposition 3-a *Main Reconfiguration Strategy in the Source CoP* (Fig. 7.1): When a member in the source CoP is reallocated to the destination CoP, her collaboration relationships are reconfigured to other remaining members with the most
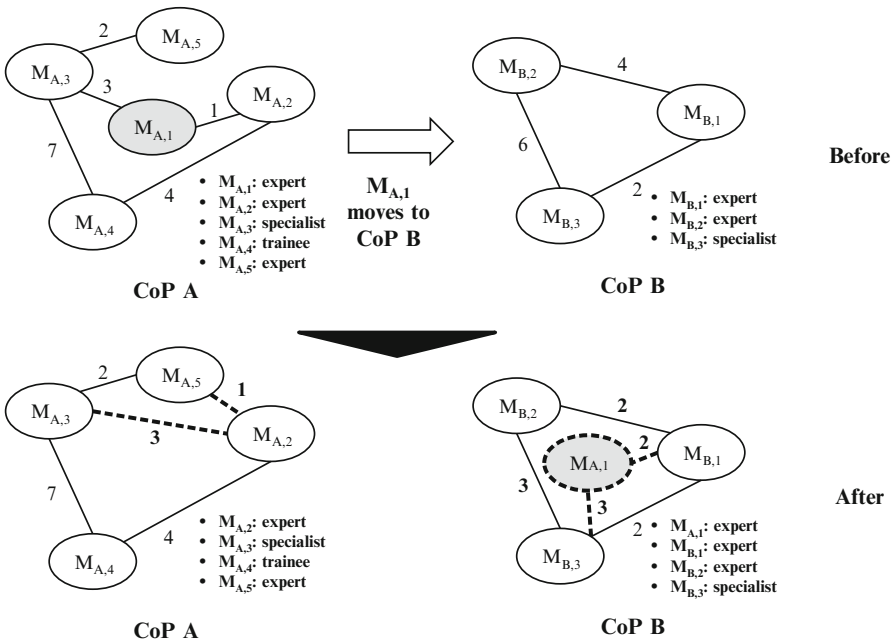
Fig. 7.1 Reconnection and reconfiguration strategy in source and destination CoPs

active involvement in knowledge sharing in the source CoP. Members with the same or higher expertise level are preferred to members with lower expertise level.

For example, in Fig. 7.1, when an expert member $M_{A,1}$ in CoP A is reassigned to CoP B, her connections with $M_{A,3}$ (weight 3) and $M_{A,2}$ (weight 1) are reassigned to $M_{A,2}$ who has the maximum knowledge-processing capacity (total weight 4) among experts in CoP A mainly because the tasks performed by departed member tend to be reassigned to the remaining members with maximum performance capacity (Kwon et al. 2007). To prevent the loss of the total amount of CoP activity, a new direct relationship between $M_{A,3}$ and $M_{A,2}$ is assumed to take a weight of 3, a larger weight of $M_{A,1}$ because $M_{A,3}$ is not an expert but a specialist. However, a new direct connection between $M_{A,5}$ and $M_{A,2}$ is created with a weight of 1 (a smaller weight because $M_{A,5}$ is also an expert like $M_{A,2}$) to restore to-be-lost direct connection between $M_{A,1}$ and $M_{A,2}$. When there are few remaining members with the same expertise level or no members with the same expertise level in the source CoP, however, it is necessary to randomly break the tie and select a member with the same expertise level or randomly select a member with lower expertise level to reconfigure connections of the rearrangement candidate. To this end, we posit the following proposition.

Proposition 3-b *Supplemental Reconfiguration Strategy in the Source CoP*: When Proposition 3-a is not applicable for any reasons, collaboration relationships of the rearrangement candidate are reconfigured to other randomly chosen remaining members with the same (preferred) or lower expertise level in the source CoP while keeping the total knowledge sharing activities constant in the source CoP.

Note that while the total number of connections in the source CoP remain at approximately same level, connections of remaining members vary as knowledge activities of the departing member are reassigned. Therefore, it is possible that the *BIS* of the source CoP increases or decreases after reconfiguration. Once the process of disconnecting and reconnecting of collaboration relationships among remaining members in the source CoP is completed, the process of making connections between the rearrangement candidate from the source CoP and members in the destination CoP begins. This process is regulated by the following proposition:

Proposition 4 *Reconnection Strategy in the Destination CoP* (Fig. 7.1): The rearrangement candidate from the source CoP takes a half of the collaboration relationships of the member with the most active knowledge sharing activities in the destination CoP.

The reconnection of collaboration relationships in the destination CoP (CoP B) based on the proposition 3 is graphically illustrated by two figures in the right of Fig. 7.1. According to this proposition, when an expert member $M_{A,1}$ from the source CoP, she takes a half of $M_{B,2}$'s collaboration relationships before rearrangement because $M_{B,2}$ performs the maximum level of knowledge sharing activities. In particular, the collaboration relationship of $M_{B,2}$ with $M_{B,3}$ has a weight of 6. Then, when $M_{A,1}$ is connected to $M_{B,3}$, the original weight of a collaboration relationship between $M_{B,2}$ and $M_{B,3}$ is reduced to a half (i.e., 3) and the lost weight is distributed to the new direct connection made between $M_{A,1}$ and $M_{B,3}$ (weight of 3). The relationship of $M_{B,2}$ with $M_{B,1}$ is also reconnected similarly. $M_{A,1}$ makes a new

connection with $M_{B,1}$ by taking a half (i.e., 2) of the original weight of a collaboration relationship between $M_{B,2}$ and $M_{B,1}$.

Note that the direction of knowledge flow between members is not specified in Fig. 7.1 mainly to avoid unnecessarily complicated presentation of our reconnection and reconfiguration propositions. However, this study explicitly considers the flow directions of collaborations (both in- and out-degree) in the process of implementing reconnection and reconfiguration strategies in the following sections to realistically model knowledge sharing processes with knowledge creation and knowledge consumption.

7.3.2 Knowledge Sharing Data Sets

To show the improvement of knowledge sharing activities among CoPs members by reconfiguring CoP memberships, we start with real CoP activity data sets from Company P that currently supports 1,600 CoPs, with a total number of CoP participants of about 89,000 employees. The data sets used in this study is a sampled data sets that contain the knowledge sharing activities of 3,730 employees (representing 4,414 members because 568 employees, or 15.2 %, engage in more than one CoP) across 59 CoPs from four departments: Iron & Steel (I01–I14), Maintenance (M01–M15), Rolling (R01–R14), and Staff (S01–S16). To obtain reliable and representative information, we sample about the same number of CoPs (between 14 and 16) and a similar number of members from each department (910 members from Staff to 1,296 members from Rolling). Each CoP has an average of about 7.48 employees, and each employee creates 4.4 messages and consumes 72.2 messages. The total number of members is 4,414 with 838 experts (e.g., executive, VP, senior managers), 1,584 specialists (e.g., junior managers), and 1,992 trainees (e.g., new employees). For further analysis, CoP members are first classified into four categories—$Member^{CO}$ (core player, 5.7 %), $Member^{CR}$ (knowledge creator, 7.7 %) and $Member^{CS}$ (knowledge consumer, 15.6 %), $Member^{IA}$ (inactive player, 71.0 %)—based on the information captured in knowledge transfer matrix. Then, 59 CoPs are classified into four types: knowledge sharing community (CoP^{SH}, 10.2 %), knowledge storing community (CoP^{ST}, 1.7 %), knowledge learning community (CoP^{LR}, 28.8 %), and inactive community (CoP^{IA}, 59.3 %). For detailed descriptions of member types, CoP types, and classification schemes, the readers are advised to refer to Lee et al. (2012).

In our study, the pervasiveness of the *ith* bottleneck is measured as the proportion of members who actually cause the *ith* bottleneck out of all members who can cause it (e.g., all experts and specialists) while the relative priorities (w_i) of the *ith* bottleneck are based on subjective assessments on the importance of each bottleneck from two experts. The derived relative normalized priorities of the bottleneck categories and types are summarized in Table 7.1.

According to Table 7.1, the knowledge drain bottleneck is twice as important as the master-apprenticeship bottleneck (0.667 vs. 0.333) in determining the capacity of the *BIS* metric. In addition, while two bottlenecks in the knowledge drain category

Table 7.1 Relative weights of six bottlenecks

Bottleneck category priority		Bottleneck type priority		Relative priority
Master-apprenticeship	0.333	Bottleneck 1	0.597	0.199
		Bottleneck 2	0.214	0.071
		Bottleneck 3	0.101	0.034
		Bottleneck 4	0.088	0.029
		Total	1.000	0.333
Knowledge drain	0.667	Bottleneck 5	0.500	0.333
		Bottleneck 6	0.500	0.333
		Total	1.000	0.667

Source: Lee et al. (2012)

(i.e., Bottlenecks 5 and 6) are equally important, Bottleneck 1 is considered most important, followed by Bottlenecks 2, 3, and 4 in the master-apprenticeship category. Finally, multiplying the bottleneck category priority by the bottleneck type priority establishes the relative priority of each bottleneck. Examining the relative priority values, we find that Bottlenecks 5 and 6 (0.333 for each) are the most important, followed by Bottlenecks 1 (0.199) and 2 (0.071). While relative weights in Table 7.1 are subjective and could be different with different decision makers' preferences, our general analysis framework is still applicable and obtained managerial insights will be useful.

Using relative weights in Table 7.1, we compute the values of *BIS* of 59 CoPs to measure how serious each (and aggregated) bottleneck is and present them in Table 7.2. One notable fact is that the values of *BIS* of CoPs in different departments and even in the same department are very different. For example, we note that CoPs in Staff department have the highest value of *BIS* on average and the CoP with the highest *BIS* in Staff department is S14 and its *BIS* value is about 3.5 times higher than that of S01 (0.559 vs. 0.155). We believe that by dynamically reconfiguring CoP memberships within the same department, the organization may evenly distribute active participants of collaboration network across CoPs and eliminate bottlenecks associated with CoP members who have to respond to so many requests from non-experts. This will ultimately improve the organizational knowledge sharing climate (Constant et al. 1996; Huber 2001; Orlikowski 1993) and employees' intention to share their knowledge (Ardichvili et al. 2003; Bock et al. 2005).

7.3.3 Simulation Experiments Setups

We adopt a computer simulation method as a principal analysis tool to test the effectiveness of the proposed system with propositions to decrease the total sum of *BIS* in the organization. Note that simulation method offers great flexibility and robustness to gain insights into the real-world situation by testing various scenarios in an artificially created and controlled environment (Starbuck 2004; Kwon et al. 2007).

Table 7.2 BIS of 59 CoPs

CoP ID	BIS	CoP ID	BIS	CoP ID	BIS	CoP ID	BIS
I01	0.245	M01	0.227	R01	0.349	S01	0.155
I02	0.395	M02	0.369	R02	0.456	S02	0.297
I03	0.227	M03	0.229	R03	0.212	S03	0.436
I04	0.356	M04	0.320	R04	0.219	S04	0.350
I05	0.248	M05	0.237	R05	0.327	S05	0.329
I06	0.259	M06	0.451	R06	0.275	S06	0.481
I07	0.274	M07	0.283	R07	0.366	S07	0.395
I08	0.322	M08	0.379	R08	0.266	S08	0.492
I09	0.274	M09	0.366	R09	0.268	S09	0.374
I10	0.241	M10	0.318	R10	0.255	S10	0.221
I11	0.258	M11	0.350	R11	0.451	S11	0.438
I12	0.283	M12	0.186	R12	0.360	S12	0.548
I13	0.362	M13	0.371	R13	0.246	S13	0.302
I14	0.342	M14	0.392	R14	0.272	S14	0.559
–	–	M15	0.229	–	–	S15	0.529
–	–	–	–	–	–	S16	0.520
Avg. BIS	0.292	Avg. BIS	0.314	Avg. BIS	0.309	Avg. BIS	0.402

```
1:    BIS_cur:=Calculate the total sum of BIS;
2:    while (iteration):
3:            CoP_source:=select source CoP;
4:            Member_move:=select Member;
5:            CoP_destination:=select destination CoP;
6:            move Member_move from CoP_source to CoP_destination;
7:            BIS_new:=calculate the new total sum of BIS;
8:            If (BIS_new < BIS_cur):   // improved
9:                    BIS_cur:=BIS_new;
10:           else:                              // not improved
11:                   Cancel the movement and restore original network;
12:           end if;
13:   end while;
```

Fig. 7.2 Pseudo code

We carry out simulations with two different strategies to determine the destination CoP in addition to general propositions in previous section. These strategies are different in the sense whether it considers the distribution of expertise levels of CoP members in the selection process of the destination CoP (Strategy B) or not (Strategy A). In other words, the ultimate goal of Strategy B is to evenly distribute experts across CoPs to catalyze an organizational knowledge sharing climate that will boost employees' willingness to share their knowledge (Ardichvili et al. 2003; Bock et al. 2005). We present the pseudo code of rearrangement process in Fig. 7.2.

7.4 Experimental Results: BIS Improvement and Network Structure

7.4.1 Comparison of BIS Improvements and CoP Types Distribution

We carry out simulations with 5,000 iterations for each strategy and measure the improvement of *BIS* (= (BIS_{old}–BIS_{new})/BIS_{old}) over iterations as CoP members are rearranged and collaboration networks are reconfigured. We graphically present such information in Fig. 7.3.

Overall, both strategies significantly improve the healthiness of collaboration networks in terms of BIS value. To our surprise, however, Strategy A results in greater improvement (18 %) than Strategy B (10 %). We partially attribute this finding to the fact that Strategy A makes it possible to assign more rearrangement candidates into destination CoPs during the fixed iteration because it does not enforce any extra eligibility requirements on destination CoPs while Strategy B searches for destination CoPs that satisfy an additional distribution requirement of experts. In particular, Strategy B does not make further improvement in terms of *BIS* values after the iteration of 3,500, indicating that there is no available destination CoPs. Interestingly, Strategy B makes steeper improvement at early iterations (up to 255 iterations) mainly because it heuristically finds better fitting destination CoPs for chosen rearrangement candidates.
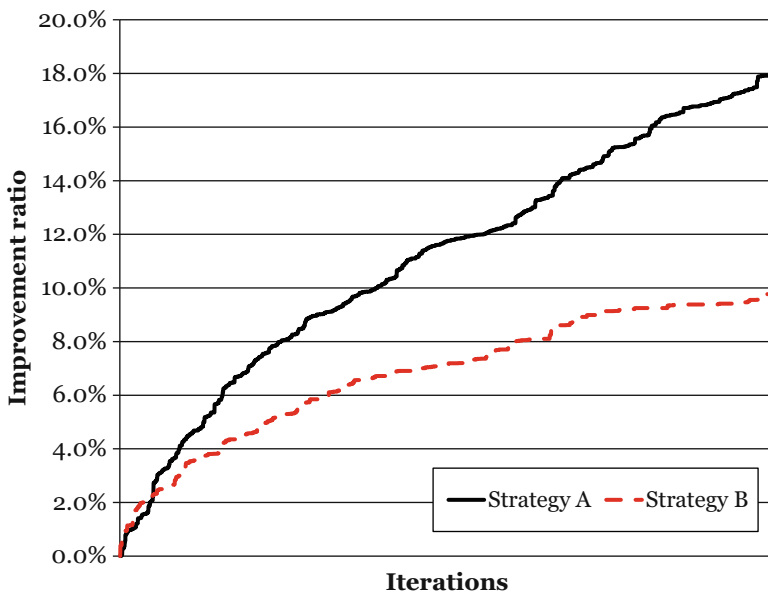
Fig. 7.3 Improvement of BIS values by strategy A & B
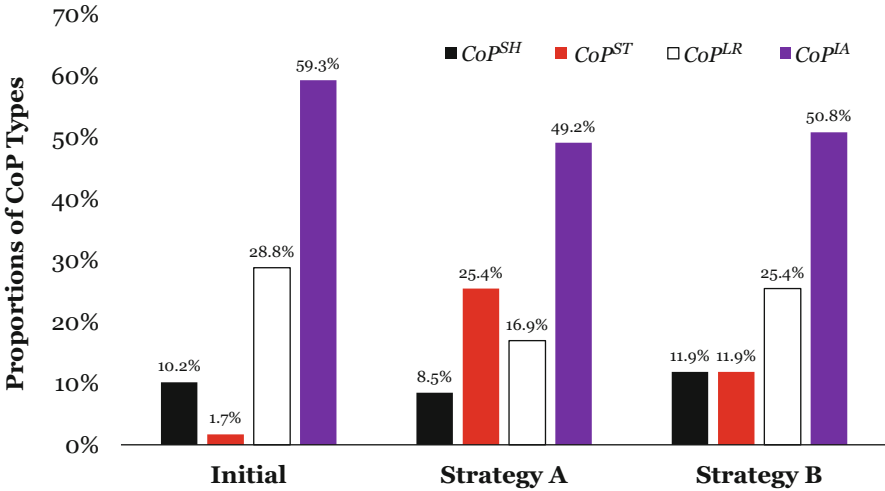
Fig. 7.4 Proportions of CoP types

The effectiveness of Strategy A and B is also compared in terms of the proportions of each CoP types. According to Fig. 7.4, both Strategy A and B significantly increase the proportion of CoP^{ST} type (from 1.69 % to 25.42 % and 11.86 %, respectively) and also significantly decrease the proportion of CoP^{IA} type (from 59.32 % to 49.15 % and 50.85 %, respectively). Therefore, both strategies greatly improve the structural healthiness of collaboration networks with the organization. One of major differences between two strategies come from the fact that Strategy A results in a much higher proportion of CoP^{ST} than (25.42 % vs. 11.86 %) Strategy B while it results in a much lower proportion of CoP^{LR} than Strategy B (16.95 % vs. 25.42 %). Therefore, we can conclude that the major improvement of BIS value via Strategy A over Strategy B is due to the significantly increased proportion of CoP^{ST} type. However, we also note that Strategy A slightly decrease the proportion of CoP^{SH}, insinuating a possible negative impact over a long-term period.

7.4.2 Improvement of Collaboration Network Structures

To illustrate the improvement of structural healthiness, we graphically show the change of the network structure of an exemplar CoP, S05 in Fig. 7.5 in which triangle, rectangle, and circle represents expert, specialist, and trainee, respectively. This CoP was one of inactive CoPs in which eight members were orphaned without connections to other members, indicating that few experts were connected with too many non-experts and many experts and specialists (two experts and five specialists) were not fully utilized. However, after the reconfiguration of its collaboration networks via either Strategy A or Strategy B, all members except two are now connected to other members and the loads of few experts with heavy loads are nicely spread out
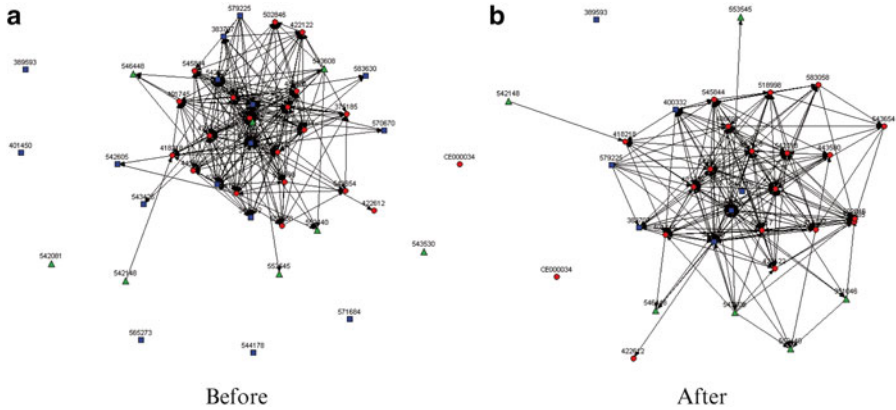
Fig. 7.5 Network structures of S05 before and after reconfiguration

to other experts or specialists. There are no experts without active connections to other members. Overall, the new collaboration network structure of S05 shows the typical pattern of *CoP^{ST}* and its *BIS* is improved by 45 % and 15 % for Strategy A (from 0.329 to 0.180) and Strategy B (from 0.329 to 0.281), respectively.

One final note in regard to the comparison of Strategy A and B is that while Strategy A results in a greater improvement in terms of *BIS* value than Strategy B, Strategy B is likely to provide a solution that reflects the desired distribution of three CoP member types. To this end, we present the proportions of member types in CoPs of Rolling department. According to Table 7.3, the proportions of experts show an extremely distorted distribution (between 1 % and 33 %) across CoPs in Strategy A induced networks, while the corresponding distribution in Strategy B induced networks is much more balanced (between 9 % and 34 %). Specialists (18–67 % vs. 21–62 %) and trainees (14–81 % vs. 16–63 %) across CoPs in Strategy B induced networks are also more evenly distributed than in Strategy A induced networks. Note that well-balanced distributions of experts and specialists are likely to arouse collaborative environments because they are ones who can create and share knowledge, and hence lead CoPs to a higher level of knowledge sharing activities from a long-term perspective. However, Strategy A is still useful in the sense that it presents an upper bound of *BIS* improvement when balanced distributions of human experts are not considered at all.

7.5   Conclusion

This study simulates two reconfiguration strategies to modify the structure of collaboration networks established among 4,414 members in 59 CoPs and reports that the proposed approach effectively decreases potential threats to knowledge sharing and improves the structural healthiness of knowledge sharing networks when it is measured by *BIS*. Specifically, the number of knowledge storing CoPs is

Table 7.3 Distributions of expertise levels of CoPs in rolling department

CoP ID	Proportions with strategy A			Proportions with strategy B		
	Expert	Specialist	Trainee	Expert	Specialist	Trainee
R01	18 %	46 %	36 %	12 %	44 %	44 %
R02	23 %	45 %	32 %	11 %	46 %	43 %
R03	20 %	40 %	40 %	13 %	38 %	48 %
R04	13 %	30 %	57 %	11 %	33 %	56 %
R05	10 %	27 %	63 %	14 %	46 %	40 %
R06	18 %	*67 %*	*14 %*	22 %	*62 %*	*16 %*
R07	19 %	49 %	32 %	*34 %*	33 %	33 %
R08	13 %	27 %	60 %	11 %	33 %	56 %
R09	*1 %*	*18 %*	*81 %*	*9 %*	36 %	55 %
R10	15 %	64 %	21 %	11 %	46 %	43 %
R11	5 %	43 %	51 %	26 %	33 %	40 %
R12	22 %	53 %	25 %	26 %	36 %	38 %
R13	8 %	38 %	54 %	11 %	42 %	48 %
R14	*33 %*	33 %	33 %	16 %	*21 %*	*63 %*

significantly increased while the number of inactive CoPs is decreased. Another important structural improvement is that expert members with an appropriate amount of collaboration burden are evenly distributed across CoPs, making it possible for each CoP to further transform into a full-fledged CoP with a sufficient number of experts from whom trainees may learn. Overall, the proposed approach helps organizations by eliminating structural bottlenecks and evenly distributing both experienced and unexperienced members.

Findings from this research contribute to knowledge sharing and management community in the sense that we quantitatively measure the seriousness of barriers to knowledge sharing activities in CoPs and demonstrate the impact of reconfiguration strategies by linking current activity data sets collected from real CoPs to simulated future outcomes. Practitioners may benefit from adopting the proposed approach not only to improve the current structural healthiness of knowledge networks in CoPs for a short-term period but also to establish a stable structure of organizational collaboration networks toward active knowledge sharing for a long-term period with no needs to change their current IT infrastructure, rewards incentives, or organizational hierarchy.

While the proposed system bears methodological contributions and presents several managerial insights, it is limited in the sense that reconfiguration strategies do not consider other important individual psychological factors or organizational cultures that may affect members' motivations to share their implicit and explicit knowledge through collaboration networks and regulate reconfiguration propositions. Therefore, in our follow-up research, we intend to extend the current research to compare the effectiveness and usefulness of the proposed approach within two completely different organizational cultures (e.g., vertical or horizontal organizational culture). Another possible future research is to develop a new methodology with reconfiguration propositions and measure the synergy effects when two or more intimate members are allowed to move at the same time.

Biography

Suchul Lee   Dr. Suchu Lee is a senior researcher in Department of Future R&D Strategy at Korea Institute of Science and Technology Information (KISTI). He received Ph.D. in Industrial and Management Engineering from the Pohang University of Science and Technology (POSTECH). Dr. Lee's primary research interests include management information systems, knowledge management, decision support system, and strategic management of technology..

Yong Seog Kim   Dr. Yong Seog Kim is an associate professor in Management Information Systems department at the Utah State University. He received his M.S. degree in Computer Science and Ph.D. in Business Administration from the University of Iowa. Dr. Kim's primary research interest is in decision support systems utilizing various data mining (KDD) algorithms such as variable selection, clustering, classification, and ensemble methods. His papers have appeared in Management Science, Decision Support Systems, Intelligent Data Analysis, Expert Systems with Application, and Journal of Computer Information Systems, and conference proceedings of KDD, AMCIS, DSI, HICSS, and many others. Dr. Kim currently serves on the editorial board of the Journal of Computer Information Systems, Journal of Information Technology Cases and Applications, and Journal of Emerging Trends in Computing and Information Sciences.

Euiho Suh   Dr. Euiho Suh is a full professor in Department of Industrial and Management Engineering at Pohang University of Science and Technology (POSTECH), Republic of Korea. He has a B.S. in Industrial Engineering from Seoul National University, Republic of Korea; two master's degrees in Industrial Engineering from the Korea Advanced Institute of Science and Technology (KAIST) and Stanford University, USA; and a Ph.D. in Management Information Systems from University of Illinois at Urbana-Champaign, USA. His research interests include management information systems, decision support systems, strategic management of information systems, information and knowledge management, and strategic management of technology. His papers appeared in Decision Support Systems, Journal of Knowledge Management, International Journal of Information Management, IEEE Transactions on Engineering Management, Expert Systems with Applications, Knowledge and Process Management, Electronic Commerce Research and Applications, and many others.

References

Ahuja, M. K., & Carley, K. M. (1999). Network structure in virtual organizations. *Organization Science, 10*(6), 741–757.

Al-Alawi, A. I., Al-Marzooqi, N. Y., & Mohammed, Y. F. (2007). Organizational culture and knowledge sharing: Critical success factors. *Journal of Knowledge Management, 11*(2), 22–42.

Ardichvili, A., Page, V., & Wentling, T. (2003). Motivation and barriers to participation in virtual knowledge-sharing communities of practice. *Journal of Knowledge Management, 7*(1), 64–77.

Bang, H. K., Ellinger, A. E., Hadjimarcou, J., & Traichal, P. A. (2000). Consumer concern, knowledge, belief, and attitude toward renewable energy: An application of the reasoned action theory. *Psychology & Marketing, 17*(6), 449–468.

Baum, J. R., Locke, E. A., & Smith, K. G. (2001). A multidimensional model of venture growth. *Academy of Management Journal, 44*(2), 292–303.

Bearden, W. O., & Etzel, M. J. (1982). Reference group influence on product and brand purchase decisions. *Journal of Consumer Research, 9*(2), 183–194.

Blau, P. M. (1964). *Exchange and power in social life*. New York: Wiley.

Blyler, M., & Coff, R. W. (2003). Dynamic capabilities, social capital, and rent appropriation: Ties that split pies. *Strategic Management Journal, 24*(7), 677–686.

Bock, G. W., Zmud, R. W., Kim, Y. G., & Lee, J. N. (2005). Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate. *MIS Quarterly, 29*(1), 87–111.

Bock, G. W., Kankanhalli, A., & Sharma, S. (2006). Are norms enough? The role of collaborative norms in promoting organizational knowledge seeking. *European Journal of Information Systems, 15*(4), 357–367.

Bock, G. W., Sabherwal, R., & Qian, Z. J. (2008). The effect of social context on the success of knowledge repository systems. *IEEE Transactions on Engineering Management, 55*(4), 536–551.

Borgatti, S. P., & Foster, P. C. (2003). The network paradigm in organizational research: A review and typology. *Journal of Management, 29*(6), 991–1013.

Bosua, R., & Scheepers, R. (2007). Towards a model to explain knowledge sharing in complex organizational environments. *Knowledge Management Research & Practice, 5*(2), 93–109.

Botkin, J. W. (1999). *Smart business: How knowledge communities can revolutionize your company*. New York: The Free Press.

Brown, J. S., & Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science, 2*(1), 40–57.

Cantner, U., & Graf, H. (2006). The network of innovators in Jena: An application of social network analysis. *Research Policy, 35*(4), 463–480.

Constant, D., Sproull, L., & Kiesler, S. (1996). The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization Science, 7*(2), 119–135.

Cross, R. L., & Parker, A. (2004). *The hidden power of social networks: Understanding how work really gets done in organizations*. Boston: Harvard Business School Press.

Cross, R., Parker, A., & Borgatti, S. (2000). A bird's-eye view: Using social network analysis to improve knowledge creation and sharing. *Knowledge Directions, 2*(1), 48–61.

Dalkir, K. (2005). *Knowledge Management in Theory and Practice*. Burlington: Elsevier Butterworth-Heinemann.

De Laat, M., Lally, V., Lipponen, L., & Simons, R.-J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. *International Journal of Computer-Supported Collaborative Learning, 2*(1), 87–103.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*(1), 105–115.

Fritz, M. B. W., Narasimhan, S., & Rhee, H.-S. (1998). Communication and coordination in the virtual office. *Journal of Management Information Systems, 14*(4), 7–28.

Fulk, J., Heino, R., Flanagin, A. J., Monge, P. R., & Bar, F. (2004). A test of the individual action model for organizational information commons. *Organization Science, 15*(5), 569–585.

Griffith, T. L., Sawyer, J. E., & Neale, M. A. (2003). Virtualness and knowledge in teams: Managing the love triangle of organizations, individuals, and information technology. *MIS Quarterly, 27*(2), 265–287.

Gupta, K. S. (2008). A comparative analysis of knowledge sharing climate. *Knowledge and Process Management, 15*(3), 186–195.

Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research, 18*(4), 323–342.

Helms, R. W. (2007). Redesigning communities of practice using knowledge network analysis. In A. Kazi, L. Wohlfart, & P. Wolf (Eds.), *Hands-on knowledge co-creation and sharing: Practical methods and techniques* (pp. 253–273). Stuttgart: Knowledge Board.

Helms, R. W., Ignacio, R., Brinkkemper, S., & Zonneveld, A. (2010). Limitations of network analysis for studying efficiency and effectiveness of knowledge sharing. *Electronic Journal of Knowledge Management, 8*(1), 53–68.

Hinds, P., & Kiesler, S. (1995). Communication across boundaries: Work, structure, and use of communication technologies in a large organization. *Organization Science, 6*(4), 373–393.

Huber, G. P. (2001). Transfer of knowledge in knowledge management systems: Unexplored issues and suggested studies. *European Journal of Information Systems, 10*(2), 72–79.

Iyengar, R., Van den Bulte, C., & Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science, 30*(2), 195–212.

Janz, B. D., & Prasarnphanich, P. (2003). Understanding the antecedents of effective knowledge management: The importance of a knowledge-centered culture. *Decision Sciences, 34*(2), 351–384.

Jones, Q., Ravid, G., & Rafaeli, S. (2004). Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information Systems Research, 15*(2), 194–210.

Kilduff, M., & Tsai, W. (2003). *Social networks and organizations*. Thousand Oaks: Sage.

Kim, Y., & Lee, B. (1995). R&D project team climate and team performance in Korea: A multidimensional approach. *R&D Management, 25*(2), 179–196.

Kim, W. C., & Mauborgne, R. (1997). Fair process: Managing in the knowledge economy. *Harvard Business Review, 75*(4), 65–75.

Kwon, D., Oh, W., & Jeon, S. (2007). Broken ties: The impact of organizational restructuring on the stability of information-processing networks. *Journal of Management Information Systems, 24*(1), 201–231.

Lee, C., & Green, R. T. (1991). Cross-cultural examination of the Fishbein behavioral intentions model. *Journal of International Business Studies, 22*(2), 289–305.

Lee, S., Suh, E., & Kim, Y. S. (2012). Health diagnosis of communities of practices (CoPs). In *Proceedings of AMCIS*, Seattle, WA.

Lesser, E. L., & Storck, J. (2001). Communities of practice and organizational performance. *IBM Systems Journal, 40*(4), 831–841.

Marwell, G., & Oliver, P. (1993). *The critical mass in collective action*. Cambridge: Cambridge University Press.

McDermott, R. (1999). Why information technology inspired but cannot deliver knowledge management. *California Management Review, 41*(4), 103–117.

McDermott, R., & Archibald, D. (2010). Harnessing your staff's informal networks. *Harvard Business Review, 88*(3), 82–89.

Molm, L. D. (1997). Risk and power use: Constraints on the use of coercion in exchange. *American Sociological Review, 62*(1), 113–133.

Moreno, J. L. (1934). *Who shall survive? A new approach to the problem of human interrelations*. Washington, DC: Nervous and Mental Disease Publishing Co.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *The Academy of Management Review, 23*(2), 242–266.

Nebus, J. (2004). Learning by networking: Knowledge search and sharing in multinational organizations. In *Proceedings of the 46th academy of international business, bridging with the other: The importance of dialogue in international business*, Stockholm.

Oh, W., & Jeon, S. (2007). Membership herding and network stability in the open source community: The ising perspective. *Management Science, 53*(7), 1086–1101.

Orlikowski, W. J. (1993). Learning from notes: Organizational issues in groupware implementation. *The Information Society, 9*(3), 237–250.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy, 6*(1), 65–78.

Starbuck, W. H. (2004). Vita contemplativa why I stopped trying to understand the real world. *Organization Studies, 25*(7), 1271–1294.

Sveiby, K.-E., & Simons, R. (2002). Collaborative climate and effectiveness of knowledge work – an empirical study. *Journal of Knowledge Management, 6*(5), 420–433.

Thorn, B. K., & Connolly, T. (1987). Discretionary data bases a theory and some experimental findings. *Communication Research, 14*(5), 512–528.

Triandis, H. C. (1972). *The analysis of subjective culture*. Oxford: Wiley-Interscience.

Tse, D. K., Lee, K., Vertinsky, I., & Wehrung, D. A. (1988). Does culture matter? A cross-cultural study of executives' choice, decisiveness, and risk adjustment in international marketing. *The Journal of Marketing, 52*(4), 81–95.

Tuten, T., & Urban, D. (1999). Specific responses to unmet expectations: The value of linking Fishbein's theory of reasoned action and Rusbult's investment model. *International Journal of Management, 16*(4), 484–489.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204.

Von Krogh, G. (1998). Care in knowledge creation. *California Management Review, 40*(3), 133–153.

Vroom, V. H. (1964). *Work and motivation*. Oxford: Wiley.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.

Wenger, E., & Snyder, W. M. (2000). Communities of practice: The organizational frontier. *Harvard Business Review, 78*(1), 139–145.

Zhang, W., & Watts, S. (2008). Online communities as communities of practice: A case study. *Journal of Knowledge Management, 12*(4), 55–71.

Chapter 8
On the Role of Ontologies in Information Extraction

Sagnika Sen, Jie Tao, and Amit V. Deokar

Abstract The ubiquity of unstructured/semi-structured data in business decision-making presents a unique challenge as data management methods developed for structured data are not directly applicable. While such non-traditional data have already become part of many organizations' product/service offerings; most data managers admit that they lack the capability to leverage such data assets to elicit meaningful information. In this context, we discuss the use of Information Extraction (IE) methodologies to aid in the decision making process that utilizes un/semi-structured data. We focus on knowledge-based IE methodologies that are particularly suitable for business domains characterized by few subject matter experts' tacit and uncodified domain knowledge. Ontologies that encapsulate and represent domain knowledge can play a key role in enabling knowledge-based IE. In this article we conduct a comprehensive review of the extant literature on Ontology-Based Information Extraction (OBIE) and articulate four different roles ontologies play in such knowledge-based IE systems. We discuss these various roles of ontologies in relation to the various IE phases and illustrate them with a case study involving IT service contracts, which is an example of a OBIE system. Finally, we discuss open research issues related to the use of ontologies, evaluation metrics, and applications of IE in decision-making.

S. Sen (✉)
Department of Business Administration, Pennsylvania State University,
30 E. Swedesford Rd., Malvern, PA 19355, USA
e-mail: sagnika.sen@psu.edu

J. Tao
College of Business and Information Systems, Dakota State University,
820 N. Washington Avenue, Madison, SD 57042, USA
e-mail: jtao16065@pluto.dsu.edu

A.V. Deokar
Sam and Irene Black School of Business, Pennsylvania State University,
5101 Jordan Road, Burke Center 268, Erie, PA 16563, USA
e-mail: amit.deokar@psu.edu

8.1   Introduction

The practice of data management is going through a significant shift in the current decade due to the proliferation of unstructured/semi-structured data. Sources of such unstructured data include, but are not limited to business policies, contracts, presentations, social media and the like. According to a recent survey, 40 % data managers from a variety of industry sectors admit that unstructured data account for more than 25 % in their organizational data stores, particularly in the technology, service, and heavy manufacture industries (Mckendrick 2012). Yet, only 15% of management and IT professionals are aware of how to manage and leverage unstructured/semi-structured data involved in their products and/or services. As an example, let's consider the financial services sector. A financial analyst needs to predict the short and long-term outlook for a security based on a large body of financial reports, SEC documents, news releases, etc. However, better tools and systems that integrate analyst's domain knowledge and expertise in extracting relevant information from un/semi-structured data are needed to achieve this.

Given that the volume of unstructured/semi-structured data will likely keep on increasing in the near future, it is imperative that information systems be enabled with the capability of finding meaningful information from such data. One potential approach is Information Extraction (IE). IE refers to the process that "isolates relevant text fragments, extracts relevant information from the fragments, and then pieces together the targeted information in a coherent framework" (Cowie and Lehnert 1996). In this paper, we focus on the knowledge-based IE approaches that are suitable for contexts where assumptions and relationships among the different domain concepts already exist in the form of domain expertise, although they may or may not exist in an explicitly structured format (e.g., the financial service example described above). By encapsulating the domain expertise in the form of domain ontologies, IE can be performed to leverage this domain expertise to discover embedded knowledge nuggets from textual documents.

In this article, the main contribution of our work is in articulating the various roles and relationships domain ontologies play in different phases of IE, and articulating a research agenda for this text analytics methodology. In that regard, we present a comprehensive review of the extant literature and provide a systemic classification. The application of these different roles in Ontology-Based Information Extraction (OBIE) is illustrated through a study dealing with IT service contracts as data. Furthermore, we identify under-researched areas and discuss potential for future research, especially in the context of enabling knowledge-driven decision support with unstructured/semi-structured data.

The rest of the paper is organized as follows. Section 8.2 discusses the definition and various phases of Information Extraction. In Sect. 8.3, we elaborate four specific roles of ontologies in the IE processes, and how they inter-operate with Information Extraction tasks. The different ontology roles are illustrated in Sect. 8.4, whereas Sect. 8.5 discusses open research opportunities. Finally, concluding remarks are provided in Sect. 8.6.

8.2 Information Extraction

Text mining research has its roots in data mining, in that system components such as preprocessing modules, pattern-discovery and visualization techniques and algorithms have high-level architectural similarities (Feldman and Sanger 2007). Text mining is unique in how these components are detailed to process unstructured natural language text. For example, preprocessing focuses on representative feature extraction from unstructured textual data. Given the broad focus of text mining on processing natural language text, this field of research leverages developments from related areas of information retrieval, information extraction, and corpus-based computational linguistics (Feldman and Sanger 2007). Information Extraction (IE) is recognized as an area in which the primary goal is to transform unstructured data stored in document collections into a more structured intermediate format, on which different types of analytic techniques (e.g., reasoning, pattern discovery, pattern matching, querying) may be applied (Weiss et al. 2010). Grishman (1997) defined IE as "the identification of instances of a particular class of events or relationships in a natural language text, and the extraction of the relevant arguments of the event or relationship."

8.2.1 Phases of Information Extraction

Figure 8.1 illustrates two widely adopted conceptualizations of IE. The left side of Fig. 8.1 shows the conceptualization developed as a result of the Message Understanding Conference (MUC-7) program, sponsored by the Defense Advanced Research Project Agency (DARPA) in the late 1990s. The center part of Fig. 8.1 shows the IE conceptualization evolved through the Automated Content Extraction (ACE) program sponsored by the National Institute of Standards and Technology (NIST) that has superseded the MUC-7 program.

The MUC-7 framework was useful in providing an initial common terminology for the main IE tasks, as depicted in Fig. 8.1. First, the named entity recognition (NE) task is concerned with extracting domain entities or phrases, particularly noun phrases. Second, the coreference (CO) task is concerned with identifying different occurrences of the same entity, including both anaphoric resolution (e.g., '*that*' referring to '*John*' in a sentence) and proper noun resolution (e.g., different spellings for the same entity). Third, template element production (TE) task is concerned with associating descriptive attributes to generic entities extracted (e.g., '*John Doe*' is a person with alias '*JD*' and descriptor '*help desk manager*'). Fourth, template relation production (TR) task involves identifying basic context-dependent relations between identified entities (e.g., '*John Doe*' is an '*employee of*' '*ABC Inc.*'). Fifth, scenario template extraction (ST) is concerned with extending information from TE and TR to express domain and task-specific entities and relations (e.g. '*new product launch*' event linking related entities).
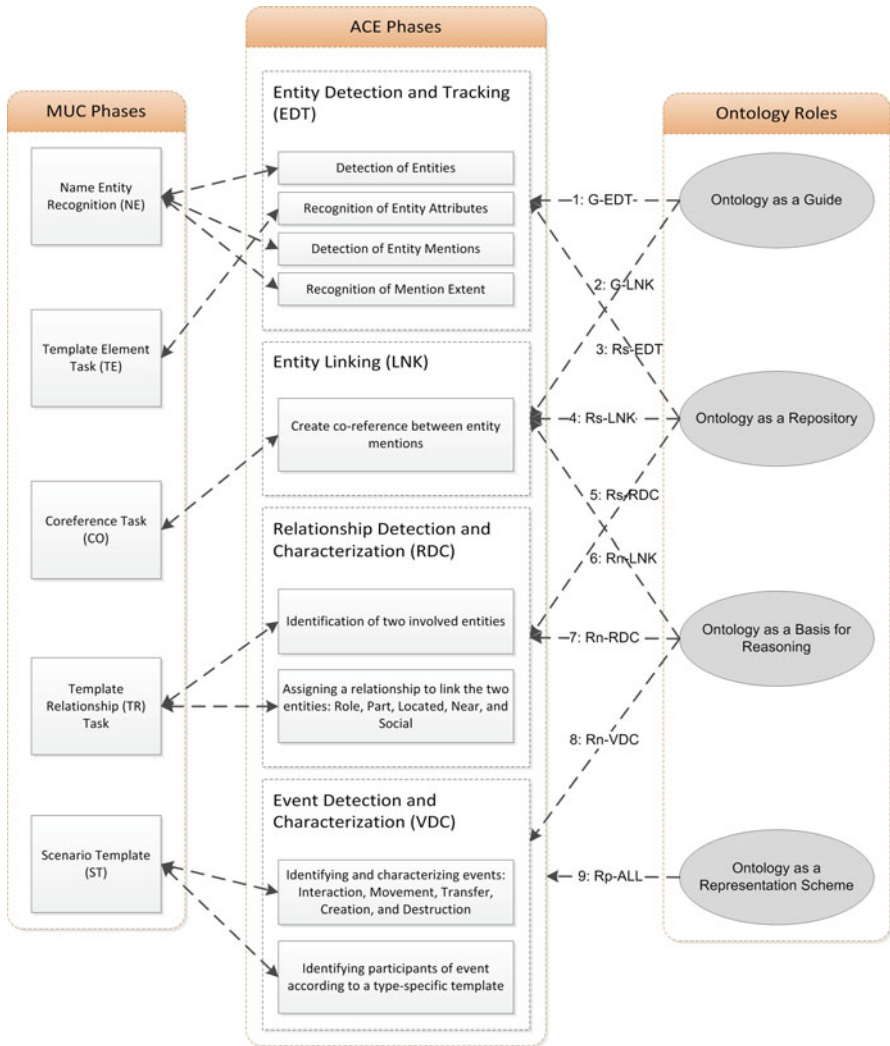
Fig. 8.1 MUC and ACE conceptualizations of information extraction phases

The ACE (2005) conceptualization details three main phases of IE: entity detection and tracking (EDT), relation detection and characterization (RDC), and event detection and characterization (VDC). The fourth phase, entity linking (LNK) was added later to this conceptualization. These IE phases build on the MUC-7 conceptualization and can be roughly mapped to it. The key difference between MUC-7 and ACE approaches is that the former is based on deriving entities from phrases mentioned in the text and associating attributes and descriptors to them to infer relations, whereas the latter is based on extracting and tracking entities through 'mention' as well as characterizing relations between these entities and events directly from text (Doddington et al. 2004).

The EDT task involves majority of the annotation work in the ACE framework, and is recognized as the pilot study in an ACE project. The four subtasks include: detection of entities, recognition of entity attributes, detection of entity mentions, and recognition of mention extent. The third and fourth subtasks are unique to the ACE approach and are based on the notion of tracking the extent of the "mention" of entities (e.g., a particular part of a sentence). The LNK task, the latest task added to the ACE framework, is focused on capturing and resolving co-references (i.e. references to the same entity), and can be considered as an extension of the CO in the MUC-7. The RDC task, added to the ACE framework in 2001–02, involves identifying relationships between the extracted entities. Once the relation links are identified, they are assigned various types (shown in Fig. 8.1) and subtypes (Turmo et al. 2006). In the VDC task, different types of events are identified and then detailed. The EDT entities serve as the participants of these events. In accordance with the scenario-specific templates, the roles of the entities and their involved attributes are defined to draw analytic inferences and meaningful relationships from the corpus.

In sum, as a more comprehensive framework compared to MUC-7, the tasks in ACE are much more detailed, and take a slightly different approach in terms of tracking entities through mention extents, and investigating events based on entities and relations grounded in the source textual documents.

The process of accomplishing these IE tasks follows two main approaches: (1) knowledge-based approaches, and (2) machine learning/statistical approaches. Knowledge-based approaches rely on a pre-defined conceptual representation of a domain of interest, utilizing such information documented by domain experts. Contrary to this, learning-based approaches use large volumes of data to elicit knowledge elements in a given domain. These approaches necessitate training data, assumed to be either readily available or created manually or semi-automatically. While both are viable approaches depending on the problem at hand, and the type and amount of available data, in this article the discussion is focused on knowledge-based approaches, and are discussed next.

8.2.2 Knowledge-Based Approaches for Information Extraction

Knowledge-based approaches utilize domain knowledge encoded in some form. Regular expressions and rule bases have been used in IE systems where domain knowledge is integrated within the IE system. In that regard, ontologies have been noted to be an excellent candidate for representing domain knowledge and have been used for supporting IE tasks. In recent years, ontology-based information extraction has emerged as a new knowledge-based approach for IE (Wimalasuriya and Dou 2010b). An OBIE system extracts particular types of knowledge from semi-structured/unstructured natural language text and provides outputs guided by ontologies.

OBIE affords several key advantages over conventional knowledge-based IE approaches. Rather than using gazetteers or lexicon lists as the semantics, the formal ontology allows semantic reasoning on extracted entities. The semantic relationships

embedded within the ontology have been proven to be the most efficient way to resolve associations/co-references among different mentions of the given concept (Feldman and Sanger 2007). Ontologies, serving as the formal and explicit representation of domain knowledge, not only support the IE tasks, but also provide a standardized format for presentation of extracted information (Wimalasuriya and Dou 2010a). Ontologies also provide a loose coupling with the IE system in that programmatic calls to modify, populate, reason on the ontologies are made from within the IE system. Given that the ontologies themselves are standalone, users can interact through ontology tools like Protégé to perform reasoning, rule execution, and such operations. As such, domain experts can update the knowledge base and the associated rule and query base independent of the core IE system.

8.3   Ontologies in Information Extraction

Grounding in extant literature related to information extraction and semantic web technologies as well as our drawing on our own experience with research projects involving information extraction, we have identified various roles that ontologies can play in information extraction. We now elaborate these different roles of ontologies in the context of IE, namely *Ontology as a Guide, Ontology as a Repository, Ontology as a Representation Scheme*, and *Ontology as a Basis for Reasoning*. The relations between the roles of ontologies and the ACE phases are shown in the right hand side of Fig. 8.1, and are discussed in detail in subsequent sections.

8.3.1   Ontology as a Guide

Given that ontologies embed domain concepts and relations, they can play a key role in annotation of text in an IE system. In particular, ontologies can help detect entities, entity attributes, and entity mentions within the EDT task (Relation 1: G-EDT), as well as for creating co-references through proper noun resolutions in the LNK task (Relation 2: G-LNK). This role of ontologies is often referred to as *semantic annotation* or *ontology-based annotation*. Ontology-based annotation is a process in which labels are added to a limited span of texts guided by external, pre-defined semantics (Maynard 2005).

Generally, two ontology-based annotation approaches have been reported in the literature: *lexicon-based* and *thesauri-based* (gazetteer lists). Both lexicons and gazetteers provide aliases for domain concepts defined in an ontology. In a lexicon-based approach, aliases for entities are provided manually in the form of properties of concepts in the ontology. For example, in an OWL ontology, annotation properties can be used to enlist aliases. This ontology can be used subsequently after using processing resources such as tokenizer, sentence splitter, part of speech tagger, and stemmer in an IE processing pipeline, so that both arbitrary sequences of lemmas

and token strings can be matched and annotated using the terms enlisted in the ontology. The GATE (General Architecture for Text Engineering) system supports such lexicon-based semantic annotation through a plugin called *Apolda* (Automated Processing of Ontologies with Lexical Denotations for Annotation) (Cunningham et al. 2002; Wartena et al. 2007). *Apolda* annotates all potential matches without disambiguation efforts, and tasks such as inconsistency handling and overlap resolution are outside the scope of this annotation (Wartena et al. 2007). For instance, an algorithm based on the traditional section ranking algorithm, namely CARROT (Cluster And Rank Related to Ontology and Thesauri), is proposed for leveraging the disambiguation needs (Malaisé et al. 2007).

In a thesauri-based approach, ontology gazetteers are created with aliases enlisted for domain concepts, which involves a simple file-based mapping between text phrases and general semantic categories. This approach is suited when there are few domain concepts with each one having a large number of possible aliases. On the contrary, a lexicon-based approach is better suited in cases where there are a large number of domain concepts with relatively few aliases for each concept. Some IE systems choose both approaches in tandem. A particular advantage of using ontologies to guide annotation tasks is that meta-level annotation allows for implicit semantic information to be discovered (Corcho 2006). A survey of semantic annotation tools is given in (Reeve and Han 2005).

8.3.2 Ontology as a Repository

Ontologies serve as a repository of domain knowledge, and can be used effectively in conjunction with IE. In particular, ontologies can further the results of ontology-based annotation to support various subtasks within the EDT, LNK and RDC tasks (Relation 3: Rs-EDT, 4: Rs-LNK and 5: Rs-RDC). This may be done by transferring the annotation information from document corpus to ontologies in the form of instances of corresponding semantic categories (classes) and related properties. This role of ontologies is referred to as *ontology population*. In terms of domain knowledge repository, ontologies may be created by domain experts. Alternatively, existing ontologies can be augmented or new ontologies built in a ground up manner by applying machine learning techniques on document corpus. This role of ontologies is referred to as *ontology learning*. Last, but not least, knowledge generated from ontology population provides the platform for subsequent VDC tasks.

Ontologies can be considered as normalized representation of text, and as such ontology population module can produce relevant instantiations of text in the form of ontology instances. The ontology instances can have various annotation properties associated with them such as start and end nodes of annotation, and document ID, which can allow entities to be tracked for subsequent reasoning. Some researchers have argued to consider IE as an ontology population task, in conjunction with ontology learning (Manine et al. 2008). Celjuska and Vargas-Vera (2004) have proposed a system for semi-automatic ontology population based on knowledge-based

IE and supervised learning mechanism. The kernel of the system is based on ontology population, although the system contains comprehensive functionality including semantic annotation and ontology learning. Few other OBIE systems based on ontology population have been reported (Alani et al. 2003; Cheng et al. 2009; Vargas-Vera et al. 2001, 2002).

Generally, ontology learning is recognized as a semi-automatic process with human intervention. In the context of IE, ontology learning refers to facilitating (semi-)automated knowledge acquisition from the unstructured sources (Maedche and Staab 2001). IE and ontology learning share a common goal of building or extracting domain-specific knowledge bases. Both ontological concepts and relations can be added or updated during the ontology learning phase. Different types of techniques have been applied for ontology learning from textual sources, and a comprehensive review can be found in Gómez-Pérez and Manzano-Macho (2005) and Buitelaar et al. (2005). Broadly, these approaches can be categorized as: linguistic-technique-based approaches, statistics-technique-based approaches, and machine-learning-based approaches. For instance, Li and Bontcheva (2007) have adopted hierarchical, supervised classification techniques from machine learning for concept learning, in that relevant tokens and their conceptual information are identified by the pre-defined classifier as candidates (of ontological classes) and then added to the ontology (as new classes) if they satisfy certain criteria. Besides adding new concepts to the ontology, mining new relationships (especially non-taxonomic relations) between ontological classes/sub-classes is recognized as another significant ontology learning task. Some research has been reported on relation learning in IE. For instance, a plug-in for mining ontological relations in a mature IE system (*Text-to-Onto*) has been proposed by Kavalec and Svaték (2005). Similar artifact is also reported by Maslennikov and Chua (2010). Further, IE systems with "ontology generators", which support ontology construction (updating classes and relations simultaneously), have been reported by Maedche et al. (2002) and Wu, Hoffmann, and Weld (2008). Related ontology construction approaches reported in other areas of study, such as a pattern ranking approach by Blomqvist and Sandkuhl (2005) can be potentially applied in IE systems.

8.3.3   Ontology as a Representation Scheme

The use of ontologies in the role of a representation scheme supports all of the tasks in the IE process. First, using ontology as a representation scheme allows decoupling of the domain knowledge (concepts, properties and their instances) from the core IE system itself. Similarly, it allows domain rule-base and queries to be loosely coupled with the IE system, and in turn allow each of these artifacts to evolve independently. Instead of interacting with different phases in IE, this particular type of role interacts with the IE process as a whole (Relation 9: Rp-ALL).

Next, ontologies in the form of a representation scheme help express domain terminologies (in the form of ontology classes) as well as related contextual infor-

mation (in the form of properties of classes and relations) in a formalized manner. This formal representation also allows domain relations to be distinguished from extraction patterns which are typically expressed as regular expressions (Feldman and Sanger 2007). The use of ontology representation scheme for annotation allows capturing detailed information not only about the entities (candidate instances) but also about the spans, mentions, and extent to the mentions. For instance, Li and Ramani (2007) have proposed an OBIE system for understanding the requirements from users in a research and design sector, in which the domain ontology is used as a representation scheme supporting different IE tasks, including semantic annotations for identifying and recognizing concepts. Also, given that domain knowledge can be represented and recognized at different levels, lexical/sub-lexical syntactic-semantic analysis can then be performed. This approach has shown to increase accuracy and decrease ambiguity (Cimiano et al. 2005). Ontologies provide a rigorous structure with model-theoretic semantics, which in turn can be used for reasoning and drawing inferences. Similarly, the representation logic embedded within ontologies allow systematic definition of logical rules and queries (Chen 2010). Use of a semantic rule-based approach can also augment ontology learning tasks, as shown by Li and Bontcheva (2007).

Last, but not least, presenting results in the form of an ontology is the distinguishing feature of OBIE, compared to traditional IE systems. Since the end users of IE systems are interested in visualizing embedded knowledge, the representation scheme determines how and to what extent the extracted knowledge can be effectively presented. Many OBIE systems use ontologies for output visualization (Maedche et al. 2002; Maedche and Staab 2000; Vargas-Vera et al. 2001; Wu and Weld 2008).

8.3.4 Ontology as a Basis for Reasoning

Ontologies can be used as a basis for reasoning and querying. This role requires use of ontologies as a representation scheme, at least in an intermediate format. It can be also argued that the advantages of reasoning and querying provide the motivations for using ontologies as a representation scheme throughout the IE process. Reasoning techniques afforded by ontologies can be classified broadly into *subsumption reasoning*, and *semantic reasoning*, which are discussed below. These reasoning techniques can support IE activities such as co-reference resolution (Relation 6:Rn-LNK), building relations (Relation 7: Rn-RDC), and event inferencing (Relation 8: Rn-VDC).

As a standardized conceptualization of the domain knowledge, the hierarchical structure of ontology provides foundation and reference for subsumption reasoning. For instance, in a document representation OBIE system reported by Manjula et al. (2003), a mathematical model based on description logic has been proposed for enriching the representation with hidden semantic relationships, in which a core inferencing tool is the subsumption reasoning mechanism. Pellet is widely applied as the reasoning engine for inferring classes and other assertions in the knowledge base of a certain domain (Sirin et al. 2007). Applications of subsumption reasoning

for other IE purposes abound (Manine et al. 2008; Maynard et al. 2005; Wimalasuriya and Dou 2010a).

Ontologies also support semantic reasoning on instances (extracted entities and relations) through the creation and execution of semantic rules using a rule engine (e.g., Jess (2008)). The rule base is codified for deriving assertions as well as adding contextual information to them in the knowledge base of a certain domain. The rules are implemented in the form of *if-then* clauses, and lead to adding new knowledge to the ontologies in the form of new relations, and properties of instances. Ontologies also serve as a basis for querying allowing knowledge workers to retrieve analytic information using a SQL-like querying language. For instance, Protégé system supports SWRL (Semantic Web Rule Language) for defining rules on OWL ontologies (Connor et al. 2005) and SQWRL (Semantic Query-enhanced Web Rule Language), for defining queries on OWL ontologies (O'Conner and Das 2009). Applications of semantic reasoning and querying in the IE context have been reported in (Buitelaar et al. 2006; Li and Bontcheva 2007; Oleneme 2009; Saggion et al. 2007).

8.3.5 An Illustration of Ontology Roles in Information Extraction

In this section, the ontology roles discussed in the above section are illustrated in the context of a research study conducted by two of the authors (Deokar and Sen 2010; Sen and Deokar 2008). The authors utilize the OBIE approach to develop an automated decision support framework to elicit the interrelationships among various process elements (e.g. activities, resources, and events) and Key Performance Indicators (KPI). The framework is instantiated in a prototype named *SLA-Miner*. The architecture of the SLA-miner, aligning with aforementioned roles of ontologies is shown in Fig. 8.2. The major components of SLA-Miner are very briefly discussed below.

In *SLA-Miner*, a unified IT service ontology captures service and process related concepts and their associated hierarchy, relations, and properties. Three main modules are incorporated in *SLA-Miner*, namely "*SLA Entity Recognition*", "*SLA Context Inference*", and "*SLA Analytics*".

The *SLA Entity Recognition* component extracts key concepts of SLAs (such as service, KPI, activity, etc.) as well as the contextual information using the IT service ontology and the real-world SLA documents as the language resources. Facilitated via the text processing environment "General Architecture for Text Engineering" (GATE), the IT service ontology serves as the guide (for semantic annotation of the entities) and the repository (for incrementally adding discovered domain knowledge to the IT service ontology through ontology instantiation/population). The second stage of *SLA-Miner*, *SLA Context Inference*, identifies typical relationships between service and process entities through the combination of domain knowledge regarding IT services and context information of discovered entities. Two levels of inference
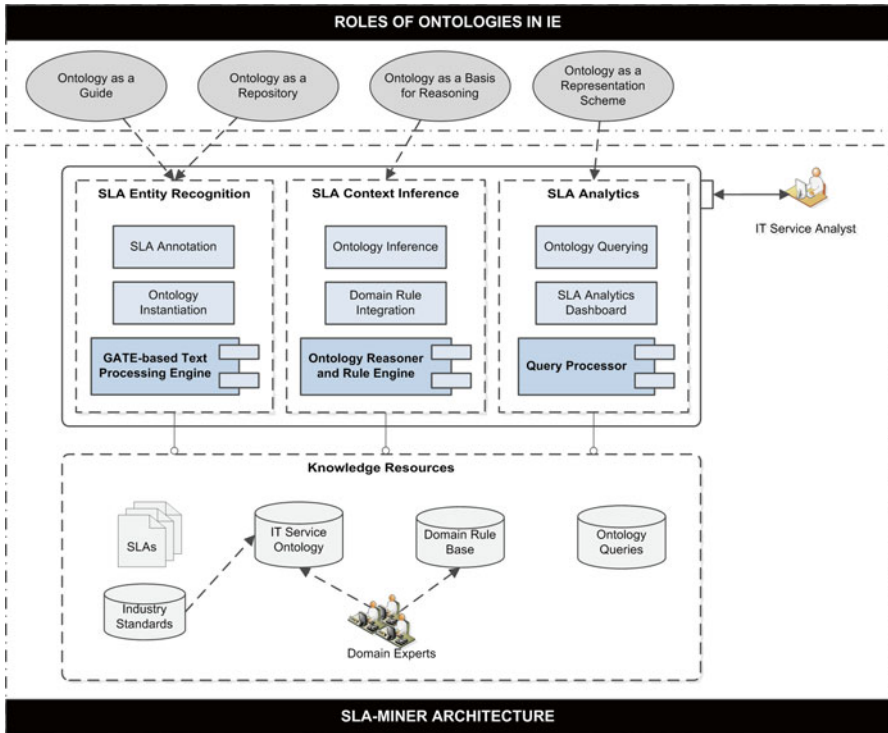
Fig. 8.2 Roles of ontologies and the architecture of SLA-miner

exist at this stage: Ontology Inference aims at obtaining knowledge assertions such as inferred classes and individuals using subsumption reasoning; Domain Rule Integration utilizes a set of logical rules (representing the domain knowledge, encoded by domain experts) to derive additional knowledge assertions (such as expected and/or unanticipated process elements and relationships) based on the IT service ontology used in conjunction with the Jess rule engine. The final stage of *SLA-Miner*, *SLA Analytics*, involves demonstrating desired information to the IT service analyst from existing and newly derived domain knowledge within the knowledge base, by executing user-defined queries in a query engine. SQWRL is used for ontology querying in *SLA-Miner* based on the Jess rule engine, which provides full-scale querying capabilities for extracting knowledge from the IT service ontology.

The efficacy of the SLA-Miner has been evaluated using IE metrics with overall precision (93 %), recall (94 %), and F-measure (93 %). The interrelationships derived were also assessed using relationship precision (66.67 %) and recall (54.55 %), which is highly comparable to the performance range (40–60 %) observed in current IE systems. The results from the feasibility and performance assessments substantiate the aforementioned four roles of ontologies in IE as guidelines for designing OBIE applications as both practicable and appropriate.

8.4   Content Analysis of Related Literature

In order to test the validity of the classification scheme presented above, we have conducted an intensive content analysis using pertinent literature that develop or apply OBIE techniques, with the proposed theoretical framework (illustrated in Fig. 8.1) as the underlying structure.

Data collection involved an extensive literature search via *Web of Science*, which provides accesses to databases such as *Science Citation Index Expanded* (SCI—Expanded), *Social Science Citation Index* (SSCI), and *Art and Humanities Citation Index* (A&HCI). To ensure the inclusiveness of our search, we have used two sets of keywords, which are "(*Information Extraction*) AND (*Ontology based*)" and "(*Text Mining*) AND (*Ontology based*)" over the time span from 2000 to 2013 with the language restricted to "English". The two searches resulted in 125 and 145 papers, respectively. By removing the duplicates, we have the initial data set of 251 papers. For analyzing the data, we have applied three different criteria for identifying the relevant literature: (a) excluding papers in which the data sources are not textual; (b) excluding papers whose focuses are not OBIE (e.g. Information Retrieval, document indexing, or non-ontology based IE); (c) excluding literature reviews, conference/task force manifestos, and summaries.

Two authors of this article reviewed the original 251 papers independently, based on aforementioned criteria. The comparison between the initial coding results indicated 87 % inter-coder agreement with respect to the relevance. After reconciliation, the authors agreed on a coding result that *172* papers out of the original 251 should be included in the content analysis.

Next, the relevant studies were categorized based on their goals of applying ontologies (with respect to the four roles of ontologies mentioned earlier). Based on the definitions of the roles in Sect. 8.3, each article is evaluated independently by two authors. Disagreements were settled through discussions and consensus. Some studies spanned across multiple categories since ontologies are used for multiple goals (e.g. ontology is used as a guide in entity recognition, and then used as a basis for reasoning). In such cases, these studies are assigned to the major categories (e.g. since the topic of the paper is to identify entities rather than reasoning, this work should be categorized in the "ontology as a guide" category rather than the "ontology as a basis for reasoning" category). Table 8.1 summarizes the methodological steps of data collection (literature search) and data analysis (coding) processes mentioned above.

Based on these results, the articles were aligned with the nine relations illustrated in Fig. 8.1. As defined in Sect. 8.3.3, we define Relation 9: Rp-ALL as the relationship between the role of "ontology as a representation scheme" and the IE process (as a whole)—thus, the articles categorized in Relation 9: Rp-ALL are the studies focusing on creating a User Interface (UI)/knowledge representation portal explicitly.

The results of these alignments are shown in Fig. 8.3. It is obvious that using ontologies as a guide for identifying/extracting entities (Relation 1: G-EDT) has been well studied (with 53 articles). It is partially because that EDT is the most mature phase in the IE process. On the contrary, Relations 2: G-LNK, 4: Rs-LNK, and 6: Rn-LNK only have 10, 2, and 3 articles respectively, probably because

Table 8.1 Summarization of literature search and coding processes

Step	Description
Step 1: Literature search	Web of science (SCI-expanded, SSCI, and A&HCI), keyword1 = "(*information extraction*) AND (*ontology based*)", keyword2 = "(*text mining*) AND (*ontology based*)", time span limited to "2000–2013", resulted in 251 unique articles
Step 2: Coding for relevant articles	Coding criteria: (a) excluding papers in which the data sources are not textual; (b) excluding the papers whose focuses are not OBIE (e.g. information retrieval, document indexing, or non-ontology based IE); (c) excluding literature reviews, conference/task force manifestos, and summaries
	Two authors coded the data set independently, with an inter-coder reliability of 87 %. With reconciliation, *172* articles were included in the content analysis with 100 % inter-coder reliability
Step 3: Coding for categorization sased on ontology roles	Categorizing selected *172* articles with respect to the four roles of ontologies. If an article spanned to multiple categories, the major one is selected (no article was assigned to more than one categories). Any disagreement was solved by consensus, which result in:
	Ontology as a guide: 63 articles
	Ontology as a repository: 44 articles
	Ontology as a representation scheme: 21 articles
	Ontology as a basis for reasoning: 44 articles
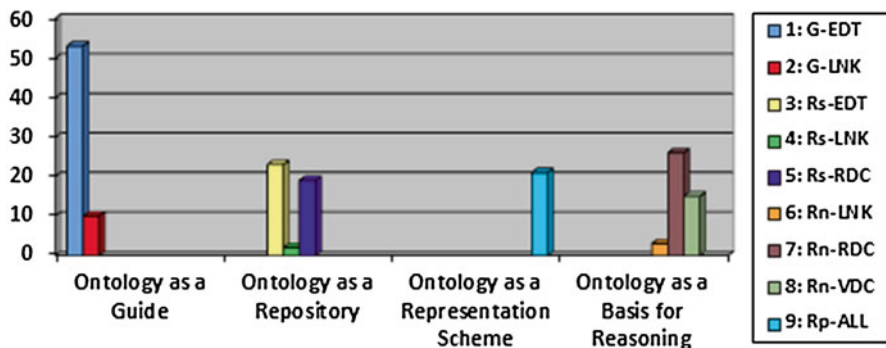
Fig. 8.3 Distribution of articles with respect to the relations illustrated in Fig.

LNK, as the newest phase in the ACE framework, requires further exploration. Based on these observations, subsequent future research issues are discussed in the following section.

8.5   Future Research Issues

Research opportunities exist with respect to the use of ontologies in IE, evaluation of OBIE systems, and innovative applications of OBIE. Key open research issues are discussed below.

8.5.1 Ontology Construction

Current OBIE systems usually rely on a single ontology. However, given that typical business scenarios involve multiple related ontologies, OBIE approaches need to consider several issues such as collaboration among ontologies, and conflict handling (Reeve and Han 2005; Wimalasuriya and Dou 2009; Wood et al. 2004). An approach called 'bridge ontology', proposed by Xu et al. (2004), for semantic annotation is an example of the type of research required in this area.

Another open research opportunity is noted in using ontologies as underlying knowledge structures for co-reference resolution purposes, as illustrated in Fig. 8.3. Current co-reference resolution methods focus on the co-occurrence/inter-reference between concepts, which ignore the semantic meanings/relations. The semantic meanings/relations can be used to identify the implicit associations between concepts—and such relations can be discovered and reflected in ontologies (Embley 2004; Snow et al. 2006).

Relations have been recognized as important as entities in IE processes. However, the central focus of current OBIE systems is on classes and instances rather than relations (Wimalasuriya and Dou 2010b). For example, the *Text-to-Onto* system relies solely on the alignment between textual co-occurrences and implicit ontological relations, and is a possible reason for less than impressive results reported (Kavalec and Svaték 2005). Advanced techniques focusing on relation-resolution in ontologies are thus needed.

8.5.2 Ontology Learning and Instantiation

Even though many OBIE systems use pre-existing "off-the-shelf" ontologies, ontology learning (OL) has become increasingly important with respect to IE process. Since defining ontologies requires domain knowledge and expertise (e.g., using particular ontology editor or language). Facilitating ontology learning would reduce these barriers through a semi-automated approach. As such, additional research is warranted on developing source/approach-independent guidelines for ontology learning, better aligning OL tools with IE methodologies, and increasing the accuracy and efficiency of OL in the context of IE in turn reducing manual interventions (Gómez-Pérez and Manzano-Macho 2005).

Another issue in this category involves expanding ontologies from the web rather than the limited text corpora. The web is a practically infinite knowledge resource with great scalability and flexibility. Even though web crawlers are incorporated with information extractors (Alani et al. 2003; Ding et al. 2006), extant (analytical) techniques are inadequate in processing data at this volume and add them to domain ontologies. Thus, there is an emerging need for the IE researchers adopting contributions and achievements from other fields (such as big data analytics and semantic web) to enable learning ontologies from the web content directly.

Another promising avenue for future work lays in extracting attributes/properties from the textual sources and adding them in association with belonging concepts in

domain ontologies. Current extraction approaches focus on the entity extraction (EDT) and relationship extraction (RDC); however, extracting attributes defining such entities/relationships have received limited attention. These attributes are particularly helpful when constructing events based on extracted entities and relationships (VDC).

8.5.3   User Interaction with Ontologies

Future research is also needed on developing intelligent user interfaces to reduce adoption barriers because of expertise needed in dealing with extracted information, ontological representations, rule bases, and queries (Gómez-Pérez and Manzano-Macho 2005). Alani et al. (2003) have proposed an automated "narrative generation" mechanism, following a story-telling based approach, for domain experts without extensive technical knowledge. However, it is limited in the context of ontology population. More approaches need to be tested for better user interaction and adoption of OBIE.

8.5.4   Evaluation of OBIE Systems

Current OBIE systems largely rely on metrics adopted from IR, such as precision and recall. However, newer metrics such as the BDM scoring proposed by Maynard et al. (2006) geared toward OBIE are needed. Also, metrics for incorporating meaningfulness of relations extracted require research attention.

For enhancing the performance of OBIE-based reasoning, research is needed for harnessing state-of-art machine-learning/Artificial Intelligence/Data mining techniques. For instance, most OBIE studies have applied classification/clustering –based techniques; yet other intelligent methods, such as Fuzzy sets, and Artificial Neural Network (ANN), have been proven to be more efficient when dealing with unstructured/semi-structured data in different domains or even different sub-fields of IE, e.g. learning-based IE (Sarawagi 2008). Application of these techniques in OBIE would further enhance its capabilities.

Other related open research issues with respect to the integration of ontologies in IE include enhancing change control of ontologies in IE process (Embley 2004), improving the performance of co-reference resolution (Ding et al. 2006), and extracting relations with anonymous classes existed (Alani et al. 2003).

8.5.5   Application of OBIE in Different Domains

OBIE can be applied in different application domains for decision support and analytics. In the finance domain, two of the authors have recently initiated a project that involves applying the OBIE approach for analyzing initial public offering (IPO)

prospectus to assist investors as well as underwriters (Tao et al. 2014). Several other applications of OBIE abound. For example, in business intelligence domain, scalability of OBIE approaches to large, frequently changing data needs to be studied. In bioinformatics domain, opportunities exist for linking concepts in research studies to experimental results. In process management, research is required in integrating developments in process modeling and mining with information extracted from process policies.

8.6 Conclusion

In this paper, we have discussed the roles of ontologies in the overall lifecycle of IE. The rise of un/semi-structured data presents a tremendous opportunity for different IE approaches. Among these, knowledge-based IE systems present a unique opportunity in various business domains since these domains are often characterized by few subject matter experts' tacit and uncodified domain knowledge. We have characterized four roles of ontologies in knowledge-based IE systems and have aligned them with extant IE approaches, phases, as well as activities. The findings are illustrated through a case study, and evaluated through a content analysis of prior related studies. Based on our analysis, we have discussed topic areas pertaining to each role mentioned above, as well as the OBIE technique, which could inform future studies in this area.

Biography

Sagnika Sen   is an Assistant Professor of Information Systems in the School of Graduate Professional Studies at Pennsylvania State University. She received her Ph.D. from Arizona State University. She has published in a number of academic journals such as Information Systems Research, Journal of Management Information Systems, Decision Support Systems, Communications of the ACM, and Human Resources Management. Her research focuses on various aspects of Business Process Management, especially performance management, incentive design, and process analytics.

Jie Tao   is a doctoral candidate at Dakota State University. He started his doctoral study since 2010, specialized in decision support and knowledge management. His research interests include: business process management, textual analytics, semantic web, and service-oriented architecture. Jie has several journal publications and published work at international conferences such as Americas Conference for Information Systems (AMCIS), International Conference for Information Systems (ICIS), and Hawaii International Conference on System Sciences (HICSS). He is also a member of the Association for Information Systems (AIS) technology committee. He is granted the AIS technology award in 2013.

Amit V. Deokar   is an Assistant Professor of Information Systems in the Sam and Irene Black School of Business at Pennsylvania State University, Erie, Pennsylvania. His recent research interests are in decision support and analytics, business process management, and collaboration processes and technologies. He has published several conference publications, journal articles, and book chapters in these areas. He holds a BE in Mechanical Engineering from V.J. Technological Institute, Mumbai, a MS in Industrial Engineering from the University of Arizona, and a PhD in Management Information Systems from the University of Arizona. He is a member of AIS, INFORMS, ACM, and AAAI.

References

ACE. (2005). http://www.ldc.upenn.edu/projects/ACE

Alani, H., Kim, S., Millard, D., Weal, M. J., Hall, W., Lewis, P., & Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *Intelligent Systems IEEE, 18*, 14–21.

Blomqvist, E., & Sandkuhl, K. (2005). Patterns in ontology engineering: Classification of ontology patterns. *ICEIS, 3*, 413–416.

Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology learning from text: An overview. In P. Buitelaar, P. Cimiano, & B. Magnini (Eds.), *Ontology learning from text methods applications and evaluation*. Amsterdam: Ios Press.

Buitelaar, P., Cimiano, P., Racioppa, S., & Siegel, M. (2006). Ontology-based information extraction with SOBA. In *Proceedings of the international conference on language resources and evaluation.* Genoa, Italy

Celjuska, D., & Vargas-Vera, M. (2004). Ontosophie: A semi-automatic system for ontology population from text. In *Proceedings international conference on natural language processing ICON 2004.* UK: Knowledge Media Institute, The Open University.

Chen, Y.-J. (2010). Development of a method for ontology-based empirical knowledge representation and reasoning. *Decision Support Systems, 50*(1), 1–20.

Cheng, H., Lu, Y.-C., & Sheu, C. (2009). An ontology-based business intelligence application in a financial knowledge management system. *Expert Systems with Applications, 36*(2), 3614–3622.

Cimiano, P., Reyle, U., & Šarić, J. (2005). Ontology-driven discourse analysis for information extraction. *Data & Knowledge Engineering, 55*(1), 59–83.

Connor, M. O., Knublauch, H., Tu, S., Grosof, B., Grosso, W., & Musen, M. (2005). Supporting rule system interoperability on the semantic web with SWRL. In *The Semantic Web: ISWC 2005* (Lecture notes in computer science, pp. 974–986).

Corcho, O. (2006). Ontology based document annotation: Trends and open research problems. *International Journal of Metadata, Semantics and Ontologies, 1*(1), 47–57.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM, 39*(1), 80–91.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: An architecture for development of robust HLT applications. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)* (pp. 168–175). Stroudsburg, PA, USA.

Deokar, A. V., & Sen, S. (2010). Ontology-based information extraction for analyzing IT service processes. In *Proceedings of the 31st international conference on information systems (ICIS '10)*, St. Louis.

Ding, Y., Embley, D. W., & Liddle, S. W. (2006). Automatic creation and simplified querying of semantic web content: An approach based on information-extraction ontologies. In *The Semantic Web: ASWC 2006* (Lecture notes in computer science, pp. 400–414).

Doddington, G., Mitchell, A., Pryzbocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The Automatic Content Extraction (ACE) program tasks, data, and evaluation. In *Proceedings of conference on language resources and evaluation (LREC 2004)*. Lisbon, Portugal.

Embley, D. W. (2004). Toward semantic understanding: An approach based on information extraction ontologies. In *Proceedings of the 15th Australasian database conference* (pp. 3–12). Darlinghurst, Australia.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Gómez-Pérez, A., & Manzano-Macho, D. (2005). An overview of methods and tools for ontology learning from texts. *The Knowledge Engineering Review, 19*(3), 187–212.

Grishman, R. (1997). Information extraction: Techniques and challenges. In *Information extraction a multidisciplinary approach to an emerging information technology* (Lecture notes in computer science). Berlin: Springer.

Kavalec, M., & Svaték, V. (2005). A study on automated relation labelling in ontology learning. In *Ontology learning from text: Methods, evaluation, and applications*. Amsterdam: Ios Press.

Li, Y., & Bontcheva, K. (2007). Hierarchical, perceptron-like learning for ontology-based information extraction. In *Proceeding WWW '07 proceedings of the 16th international conference on World Wide Web* (pp. 777–786). ACM.

Li, Z., & Ramani, K. (2007). Ontology-based design information extraction and retrieval. *AI EDAM: Artificial Intelligence for Engineering Design, Analysis, and Manufacturing, 21*(2), 137–154.

Maedche, A., & Staab, S. (2000). Mining ontologies from text. In *Knowledge engineering and knowledge management. Methods, models, and tools: 12th international conference, EKAW 2000*, Juan-les-Pins, 2–6 Oct 2000. Proceedings, pp. 169–189.

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems, 16*(2), 72–79.

Maedche, A., Neumann, G., & Staab, S. (2002). Bootstrapping an ontology-based information extraction system. In *Intelligent exploration of the web*. Heidelberg: Springer.

Malaisé, V., Gazendam, L., & Brugman, H. (2007). Disambiguating automatic semantic annotation based on a thesaurus structure. In *Proceedings of the Conference Traitement Automatique des Langues Naturelles (TALN '07)*. Toulouse, France.

Manine, A.-P., Alphonse, E., & Bessières, P. (2008). Information extraction as an ontology population task and its application to genic interactions. In *2008 20th IEEE international conference on tools with artificial intelligence* (pp. 74–81). IEEE.

Manjula, D., Aghila, G., & Geetha, T. V. (2003). Document knowledge representation using description logics for information extraction and querying. In *Proceedings ITCC 2003. International conference on information technology: Coding and computing* (pp. 189–193). Las Vegas, NV, USA.

Maslennikov, M., & Chua, T.-S. (2010). Combining relations for information extraction from free text. *ACM Transactions on Information Systems, 28*(3), 1–35.

Maynard, D. (2005). Benchmarking ontology-based annotation tools for the Semantic Web. In *Proceedings of UK e-science programme all hands meeting (AHM2005) workshop: Text mining, e-research and grid-enabled language technology*. Nottingham, UK.

Maynard, D., Yankova, M., Kourakis, A., & Kokossis, A. (2005). Ontology-based information extraction for market monitoring and technology watch. In *Proceedings of the ESWC workshop: End user aspects of the semantic web*. Heraklion, Crete.

Maynard, D., Peters, W., & Li, Y. (2006). Metrics for evaluation of ontology based information extraction. In *WWW 2006 workshop on evaluation of ontologies for the web*. Edinburgh, Scotland.

Mckendrick, J. (2012). Big data is real and it is here: 2012 survey on managing big and unstructured data. *Technical report* (pp. 1–50). MarkLogic Corporation, San Carlos, CA, USA.

O'Conner, M., & Das, A. (2009). SQWRL: A query language for OWL. In *Proceedings of OWLED 2009 OWL: Experiences and directions. Sixth international workshop* (pp. 3–10). Chantilly, Virginia, USA.

Oleneme, D. U. (2009). Information extraction: extraction of entities, relations, events and facts from bankruptcy newswire corpora. School of Computer Science, The University of Manchester, UK.

Reeve, L., & Han, H. (2005). Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on applied computing: SAC '05* (pp. 1634–1638). New York, NY, USA: ACM.

Saggion, H., Funk, A., Maynard, D., Bontcheva, K., Court, R., & Street, P. (2007). Ontology-based information extraction for business intelligence. In *Proceeding of ISWC'07/ASWC'07 proceedings of the 6th international the semantic web and 2nd Asian conference on Asian semantic web conference* (pp. 843–856). Busan, Korea.

Sandia-National-Laboratories. (2008). Jess, the rule engine for the Java platform. *Technical report.* Albuquerque, NM, USA

Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases, 1*(3), 261–377.

Sen, S., & Deokar, A. V. (2008). Information buried in B2B contracts: Towards identifying interdependencies in IT service processes. In *Proceedings of the 14th Americas conference on information systems (AMCIS '08)*. Toronto.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. In *Web semantics: Science, services and agents on the World Wide Web* (pp. 51–53).

Snow, R., Jurafsky, D., & Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *ACL-44 proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 801–808). ACM.

Tao, J., Deokar, A. V., & El-Gayar, O. F. (2014). An ontology-based information extraction (OBIE) framework for analyzing initial public offering (IPO) prospectus. In *Proceedings of the 47th annual Hawaii international conference on system sciences (HICSS-47 '14)*. IEEE.

Turmo, J., Ageno, A., & Català, N. (2006). Adaptive information extraction. *ACM Computing Surveys, 38*(2), 4.

Vargas-Vera, M., Domingue, J., Motta, E., Shum, S. B., & Lanzoni, M. (2001). Knowledge extraction by using an ontology-based annotation tool. In *Proceedings of the workshop knowledge markup and semantic annotation (K-CAP '01)*. ACM.

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., & Ciravegna, F. (2002). MnM: Ontology driven semi-automatic and automatic support for semantic markup. In *Knowledge engineering and knowledge management. Methods, models, and tools: Ontologies and the semantic web.* Springer.

Wartena, C., Brussee, R., Gazendam, L., & Huijsen, W.-O. (2007). Apolda: A practical tool for semantic annotation. In *Proceedings of the 18th international conference on database and expert systems applications (DEXA '07)* (pp. 288–292).

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of predictive text mining*. London: Springer.

Wimalasuriya, D. C., & Dou, D. (2009). Using multiple ontologies in information extraction. In *Proceeding of the 18th ACM conference on information and knowledge management: CIKM '09* (pp. 235–244). Hong Kong, China.

Wimalasuriya, D. C., & Dou, D. (2010a). Components for information extraction: Ontology-based information extractors and generic platforms. In *Proceedings of the 19th ACM international conference on information and knowledge management (CIKM '10)* (pp. 9–18). Toronto, Canada.

Wimalasuriya, D. C., & Dou, D. (2010b). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science, 36*(3), 306–323.

Wood, M. M., Lydon, S. J., Tablan, V., Maynard, D., & Cunningham, H. (2005). Populating a database from parallel texts using ontology-based information extraction. In *9th international conference on applications of natural language to information systems, NLDB 2004* (pp. 254–264). Berlin: Springer.

Wu, F., & Weld, D. S. (2008). Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th international conference on World Wide Web* (pp. 635–644). New York, NY, USA: ACM.

Wu, F., Hoffmann, R., & Weld, D. S. (2008). Information extraction from Wikipedia: Moving down the long tail. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 731–739). New York, NY, USA: ACM.

Xu, B., Wang, P., Lu, J., Li, Y., & Kang, D. (2004). Bridge ontology and its role in semantic annotation. In *Proceedings of the international conference on cyberworlds (CW '04)* (pp. 329–334).

Chapter 9
A Quantitative Approach to Identify Synergistic IT Portfolios

Ken Pinaire and Surendra Sarnikar

Abstract Healthcare organizations continue to make large investments in health information technology to improve quality of care and lower costs. Therefore, there is an ever-growing need to have an ever-clearer understanding of how IT investments impact these organizations. Past studies have explored the impact of individual technologies or aggregate all technologies based on overall investment, but do not explore the impact of specific portfolios of information technology and their synergistic effects on healthcare quality. Based on the past studies on portfolio theory, we introduce an approach, utilizing data mining techniques and logistical regression, to identify such optimal portfolios, and explore the presence of such synergistic effects among the components of the portfolio. This multi-step approach is then applied to publically-available datasets, and the resulting candidate IT portfolios are presented. Statistical analysis is then used to test these results and demonstrate the feasibility of this approach.

9.1   Introduction

As the average age of the population in the United States increases, the demand for healthcare to treat them also rises. To support this growing need for healthcare, organizations seek ways to meet this need while simultaneously improving their performance. To this end, healthcare organizations continue to make large investments in health information technology to improve quality of care and lower costs (Monegain 2009; Pizzi 2007). Given the large investments, wide variety of technologies, and the critical

K. Pinaire (✉)
Dakota State University, 300 Gateway Centre Pkwy, Richmond, VA 23235, USA
e-mail: kpinaire@pluto.dsu.edu

S. Sarnikar
College of Business and Information Systems, Dakota State University,
820 N. Washington Ave., East Hall 7, Madison, SD 57042, USA
e-mail: surendra.sarnikar@dsu.edu

nature of healthcare, there is clearly a need for a more thorough understanding of the impact of health information technology on healthcare. Specifically, it is important to evaluate and identify how specific technologies or a combination of technologies designed to support patient care impact the quality of care services and health outcomes.

Although there have been several recent studies on the impact of IT on quality (Jamal et al. 2009; Piontek et al. 2010), conflicting findings on the impact of HIT on quality (Swanson Kazley and Diana 2011), and the narrow technology focus of many studies (Wakefield et al. 2010), has left the nature of the relationship between HIT and quality unclear. Most studies investigate individual information technologies or aggregate all information technologies into broad functional clusters without comparing the effect of specific combinations of technologies and their synergistic effects on quality of care. In this chapter, we introduce and outline a new approach to identify combinations of IT systems that demonstrate synergistic effects and exhibit positive effects on healthcare quality. We then this test this proposed approach using publically available datasets and detail the results.

9.2   Literature Review

Healthcare and Information Technology are two examples of domains which contain a wealth of empirical research. Each broadly defined domain offers an abundance of research, and while the subset of research encompassing both healthcare *and* IT is smaller, it is still extensive. In order to more fully understand the relationship between information technology and healthcare quality, in this research we explore several questions. What affect does healthcare IT have on quality? Are there specific technologies or a combination of technologies that lead to a positive impact on quality of care, and if there are, what systems are involved? To develop a response to these questions, we began with a systematic review of relevant literature.

The literature search strategy involved executing a search on the PUBMED database (www.ncbi.nlm.nih.gov/pubmed) seeking English language articles published between January 1, 2000 and June 30, 2013. While the list of articles identified through the search process is not exhaustive, it is a fair representation of recent domain literature, and provides a cross-section of not only technologies frequently studied but also of common implementation environments.

Each article was reviewed and where available details about study attributes were recorded. This included the specific HIT system, disease conditions under study, research methodology, extent of user base, context of technology use and adoption, outcome measures and facilitators and barriers to this adoption.

In terms of technologies, Electronic Medical Records (EMR) (DesRoches et al. 2010), Computerized Physician Order Entry's (CPOE) (Swanson Kazley and Diana 2011), and Clinical Decision Support Systems (CDSS) (Romano and Stafford 2011) were the most commonly investigated technologies. Systems examined were in operation at multiple facilities and in various departments.

Most studies use one of two units of analysis when determining if benefits had been realized after implementation. Half of the studies focused on the facility by

comparing a facility's performance measure (e.g. mortality rate) (Piontek et al. 2010) pre and post implementation of an HIT system to judge results. The remaining half of the studies used the patient as the unit of analysis (e.g. glucose levels, blood pressure) (Hooper et al. 2013; Hunt et al. 2009) to determine impacts. In almost all cases, longitudinal data was required to ensure temporality and thus support the author's causality conclusions.

As is evident in Table 9.1, there is no clear consensus regarding a positive or a negative impact of information technology on healthcare quality. While many studies offered strong support for the implementation of HIT (Nirel et al. 2011; Wakefield et al. 2010), almost as many found either marginal benefits (Romano and Stafford 2011; Swanson Kazley and Diana 2011), improvements to quality from some IT systems and not others (DesRoches et al. 2010), or benefits for only some patients (Loiselle et al. 2010).

Some of the papers reviewed were of particular interest as they offered specific insights from their study's perspective, but when taken in totality offer clear trends in research results. For instance, Piontek et al. (2010) pointed out that medical errors and undesirable outcomes are costly. Therefore, as facilities' severity-adjusted mortality rates declined due to the implementation of an adverse-drug-event (ADE) alert system, so did pharmacy and variable drug costs. Additionally, in the process of protecting

Table 9.1 Summary of empirical research on impact of IT on healthcare quality

Technology	Positive	Neutral	Negative	Inconclusive
EMR	Adams et al. (2003), Campbell et al. (2008), El-Kareh et al. (2009), Gaylin et al. (2011), Gluck (2009), Hunt et al. (2009), Nirel et al. (2009), Cochran et al. (2011), Hazelhurst et al. (2012), Restuccia et al. (2012)	DesRoches et al. (2010), McCullough et al. (2010), O'Connor et al. (2005), Romano and Stafford (2011), Pillemer et al. (2011)	Kern et al. (2009), Morin et al. (2009)	Bardach et al. (2009)
CPOE		McCullough et al. (2010), Swanson Kazley and Diana (2011)	Koppel et al. (2005), Roberts et al. (2010)	
CDSS	DesRoches et al. (2010), Fraenkel et al. (2003), Jean-Jacues et al. (2011), Shelley et al. (2011)	Romano and Stafford (2011)	Roberts et al. (2010)	
Other	Golob et al. (2008), Davis and Pavur (2011), Menachemi et al. (2008), Piontek et al. (2010), Yu and Houston (2007), Spielberg et al. (2011), Lucero et al. (2011), Sharkey et al. (2013), Virga et al. (2012), Restuccia et al. (2012), Frimpong et al. (2013), Cohen et al. (2013), Hooper et al. (2013)	Davis and Pavur (2011)	Furukawa and Adam (2008), Loiselle et al. (2010), Gluck (2009)	Savage et al. (2010), Deily et al. (2013), Campion et al. (2013)

patient health, a peripheral benefit of HIT systems may be physician education. Roberts et al. (2010) reveal the number of true positive alerts from an ADE alert system declined over time. This may indicate that the alerts caught by the system informed prescribers who in turn became more familiar with drug interactions; thereby reducing the occurrence of prescription errors. Contrarily, Savage et al. (2010) warns that the more complex an ePharmacy (and by extension, any HIT system) is, the more opportunity exists for the introduction of errors. When healthcare providers begin to rely entirely on the computerized system to make decisions regarding dosage, drug interaction and discharge orders, oversights can occur. These errors are almost always a result of incomplete or inaccurate information entered on the patient's behalf.

When looking at results from multiple studies, it still appears that HIT's impact on the quality of healthcare is ambiguous at best. However, there are growing signs that HIT is maturing, and benefits are beginning to be realized more reliably. Nine of the ten most recent articles included in this research reported favorable results compared to only five of the ten oldest articles appearing in this review. While only an antidotal observation, this may warrant further research as systems become more comprehensive, user friendly, and interoperable. What is clearer is that there are mitigating aspects affecting the impact of these technologies, and in some cases these dynamics are impeding their potential benefits. A more thorough understanding needs to be developed of these factors through an in-depth examination of dependent and independent variables.

9.3 Methods

In order to address this gap in research, domain literature was once again reviewed to identify an appropriate model or theoretical framework on which to base the research. This review looked at evaluations of information technology investments as well as examples of evaluations of capital investments in healthcare. Where possible, articles were specifically sought that combined both domains by reporting on evaluations of IT investments *in* healthcare.

The resulting review revealed recent articles that evaluated investments in the business (Bendoly et al. 2009; Menon et al. 2000) and healthcare (Ancker et al. 2012; Chatterjee et al. 2009; Cresswell et al. 2010; Hennington and Janz 2007; Menon et al. 2000; Myung Ko and Osei-Bryson 2004; Thrasher et al. 2006; Valdmanis et al. 2008) domains.

9.3.1 Theory Development

Several theories have been used to evaluate impact of IT on performance. However, it appears that no single theory dominates IT investment evaluation. The authors of the reviewed articles were guided by numerous models, theories and frameworks. Select papers used mature and well accepted tools such as DeLone and McLean's IT Success model (Chatterjee et al. 2009), the UTAUT model (Hennington and Janz 2007), and Data Envelopment Analysis (Bendoly et al. 2009; Valdmanis et al.

2008). Others more recently proposed models include the Actor-Network model (Cresswell et al. 2010), and Triangle model (Ancker et al. 2012).

While the studies above contribute significantly to help develop an understanding of the impact of information technology on healthcare, many of the studies either consider the impact of specific information technologies in isolation, focus on productivity and financial metrics, or aggregate several technologies into functional clusters to investigate their impact on hospital performance. Specifically, there is a no literature that explores specific combination of technologies and the synergies between various information technologies and their impact on quality of care. Portfolio theory is a potential theoretical framework that can help investigate the impact of synergies between information technologies. Portfolio theory suggests that a collection of diverse resources are used to minimize risk and maximize business opportunities (Lin et al. 2006). In order to understand the impact of portfolios and synergies, we identified and evaluated a subset of articles that explored the impact of technology portfolios on organizational performance (Table 9.2). These articles are particularly relevant as they provide a precedent for using the portfolio theory in the analysis of both information technology and healthcare investments. Lin, et al. indicate, "…a synergistic effect is expected so that the value of a technology portfolio can add up to more than the sum of its separate parts" (2006). Furthermore, Bridges, et al. (2002) confirm that portfolio theory is an appropriate choice for simultaneous analysis of multiple healthcare investments. The complimentary effects of IT systems are well supported by Zhu's (2004) examination of firms' technology infrastructure and e-commerce capabilities, Thrasher et al. (2010) research into the synergies realized from integration of multiple healthcare alliance networks, and Setia et al. (2011) analysis of how the assimilation of IT applications affect the financial performance of healthcare organizations.

Based on the past research that indicates that the complementarity and the synergistic effects between technologies in a portfolio is a key factor in influencing performance, in this study we sought to identify such optimal portfolios that positively

Table 9.2 Studies evaluating IT and healthcare investments using portfolio concept

| Study	Context of study	Guiding theoretical framework	Constructs and measures	Data and method
Bridges et al. (2002)	Multiple interventions to standardize returns	Portfolio theory	Synergy between health investments	Cost effectiveness analysis
Lin et al. (2006)	Identify is patent diversity reduces risk	Technology portfolio strategy	Synergy from IT portfolio	US Patent applications
Zhu (2004)	114 companies using e-commerce	Resource-based theory	Complementary IS	Inventory of IT, financial records
Setia et al. (2011)	IT application assimilation and use	IT portfolio theory	IT systems and net income	HIMSS & California OSHPD
Thrasher et al. (2010)	Health alliance networks	Thompson's interdependence theory	Complementary IS	Financial and quality results performance
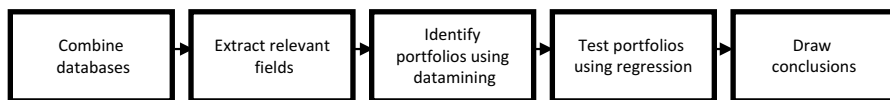
Fig. 9.1 Research approach used to identify portfolios

influence patient-outcome quality at healthcare organizations. The interactions between IT systems and the varying levels of synergies they provide may help explain the discrepancies reported by the authors of studies discussed earlier.

Specifically, this research was guided by the following two objectives:

Research Objective 1
Identify optimal portfolios of information technology that are positively associated with above average quality performance at healthcare organizations

Research Objective 2
Identify if synergistic effects exist between the components of the optimal IT portfolios. Specifically, are individual technologies more positively associated with quality when used in conjunction with other technologies within an IT portfolio than when used in isolation?

To build on the portfolio theory applications outlined in the previous section, we tested the approach depicted in Fig. 9.1.

The process begins with locating and securing suitable datasets. A suitable dataset would contain data on multiple healthcare providers, the information technology systems and applications they use, and the results of one or more healthcare quality metrics. For this research multiple datasets had to be combined in order to meet these requirements. Once the dataset was successfully merged, the required fields were identified and imported into a new database for ease of manipulation.

The third step involved applying data mining techniques through the use of decision tree classifiers to highlight candidate portfolios. These candidate portfolios where then subjected to statistical analysis to reveal intersystem synergies. Finally, optimal portfolios were identified and discussed. The rest of Sect. 3 details the process in each of these steps.

9.3.2 Data Collection, Merge and Manipulation

In order to address the research question regarding HIT's impact on quality, an analysis of hospital IT adoption records in conjunction with hospital quality results was completed. Specifically, a multi-source approach to data collection was used incorporating the 2009 HIMSS Analytics database,[1] and the 2010 version of the Medicare.gov Hospital compare database.

---

[1] The Dorenfest Institute for H.I.T. Research and Education, HIMSS Foundation, Chicago, Illinois, 2010.

9.3.3   Independent Variables

This data source identifies the IT applications in use (independent variables) for more than 34,000 healthcare facilities within the United States. These applications were clustered into clinical (e.g. electronic medical records (EMR) and picture archiving and communication system (PACS)) and business/strategic groupings (e.g. general ledger and payroll). These clusters were adapted from Setia et al. (2011) and Bhattacherjee et al. (2007) with the additional classifications of recently introduced applications. The clustering HIT in this manner has been extensively validated, and is commonly used by researchers in this field (Burke and Menachemi 2004; Burke et al. 2002; Dorenfest 2000; Menachemi et al. 2006; Pare and Sicotte 2001). The Clinical HIT cluster included applications designed to improve patient care. Because the direct impact of IT systems on healthcare outcomes was sought, only these clinical systems were used in the analysis.

9.3.4   Dependent Variables

The Medicare's Hospital Compare database provided quality measures representing patient results (dependent variables) for 4,726 facilities nationwide. The readmissions, complications and deaths details were utilized for this research. Within this portion of the data, healthcare facilities were rated with one of three ordinal values, *above the national average, equal to the national average* or *below the national average* for each of six quality measures (Heart Attack Mortality, Heart Attack Readmission, Heart Failure Mortality, Heart Failure, Readmission, Pneumonia 30 Day, and Pneumonia Readmission). The resulting combined dataset contained 3113 facilities with live and operational systems.

9.3.5   Portfolio Definition

Because synergies might be found between any two IT systems, the IT portfolio construct was defined as a combination of two or more clinical IT applications (independent variables). Many applications are present within a given facility. The applications under consideration in this research are currently implemented within healthcare facilities as reported by each facility's chosen healthcare administrator.

Step 1: Data Mining to Narrow the Search Space

The 57 predictor variables (IT applications) offered a large number of possible combinations. To deal with the volume of permutations, as well as the facility heterogeneity, data mining techniques were used to narrow the search space.

Step 2: Regression Analysis to Identify Synergies

In the next phase of the approach, the recommended portfolios were subjected to an ordinal logistic regression analysis for testing the synergistic effects among the

portfolio components. This was accomplished using statistical computing and graphics software in conjunction with the original unbalanced dataset.

The generalized form of the ordinal logistic regression model used to test the portfolios is given by:

$$y = b_0 + \sum_{i=1}^{n} b_i x_i + b_{n+1} \prod_{i=1}^{n} x_i + \sum_{k=1}^{4} b_{n+1+k} z_k$$

where b_0 is the constant and $\sum_{i=1}^{n} b_i x_i$ is the term for the predictor variables when working in isolation. $b_{n+1} \prod_{i=1}^{n} x_i$ represents the interaction of the systems, and identifies the coefficient of the portfolio. The term $\sum_{k=1}^{4} b_{n+1+k} z_k$ represents the four control variables used, and *y* represents the quality outcome (dependent variable) with values *above the national average*, *equal to the national average* or *below the national average*.

To identify if synergies exist, an interaction term was developed. While there are many ways to construct such a term, the most usual, and simplest, is to multiply the independent variables that may be involved by each other, and add that term to the equation (Flom and Strauss 2003; Harrell 2001).

9.3.6   Control Variables

Because the focus of this research is to analyze the quality performance of hospitals, other healthcare related factors that might impact this performance were controlled. Past literature has held that a facility's quality performance is likely to be influenced by size, type, ownership and case mix (Friesner et al. 2007; Setia et al. 2011). Therefore, these confounding variables were controlled using additional variables found in the Medicare dataset. The number of beds was used to control for facility size. Facilities were identified and grouped in one of three categories (General Medical, Specialty, and Critical Access). The resulting classification was used to control for facility type. Ownership data provided was identified in the dataset as—government, nonprofit and proprietary. To control for case mix, the facilities' case mix index (CMI) was retrieved from a third data source also provided by the Centers for Medicare and Medicaid Services. The provider number was once again used to match the index value to the facility.

9.4   Results

The data mining process produced extensive decisions trees for each of the quality metrics with each branch representing a different portfolio. Each branch (portfolio) was associated with one of the quality ratings (above average, average, below

average). Those portfolios associated with *above average* results were extracted and identified as recommended portfolios. The data mining process identified multiple recommended portfolios for each quality outcome.

9.4.1   Intersystem Synergies Identified

To test for the existence of synergies, a logistic regression was run on each of the recommended portfolios reported in Stage One (data mining). Each regression result returned a list of component systems and their corresponding coefficient. Additionally, the regression results identify the coefficient for combinations of systems. The coefficient indicates to what extent that system (or combination of systems) is correlated with *above the national average* results. Intersystem synergies were identified where the coefficient for the portfolio was greater than the sum of the component coefficients (Flom and Strauss 2003). Each synergistic effect identified was significant at a level of .05. The following synergies were identified within the regression results:

Heart attack mortality	
	Estimate
EDIS	−1.316550
Pharmacy_Management	−0.665235
Dictation	−1.661559
CDSS	−0.213957
Radiology_MRI	0.044432
Cardiology_IS	−3.707143
CDSS:Cardiology_IS	4.804806
Radiology_MRI:Cardiology_IS	4.468318
EDIS:Pharmacy_ Management:Dictation	1.088635

Portfolio HAM-1

Clinical Decision Support System, Cardiology Information System
Facilities with this portfolio: 847

A clinical decision support system (CDSS) is an application that healthcare providers use to analyze data in the process of making clinical decisions. A CDSS is an adaptation of the decision support system used most commonly to support business and management decision making. The cardiology information system (CIS) allows physicians to access their patients' cardiac histories as well as results when and where they are needed. Both of these systems allow physicians to access data remotely thereby offering the opportunity to consult with colleagues at different facilities in real-time. These features may account for the synergistic effect identified.

Portfolio HAM-2

Radiology: Medical Resonance Imaging, Cardiology Information System
Facilities with this portfolio: 834

The noninvasive medical diagnostic technique known as Medical Resonance Imaging (MRI), analyses the body's absorption of high-frequency radio waves. This technique is commonly used for diagnosis and treatment of cancer. An MRI system may enhance the performance of a CIS by providing the necessary imaging to monitor and manage pacemaker recipients. Cardiovascular imaging is used extensively during pacemaker implantation and is a required component to offer the coronary angiography service. The tight coupling of these two systems may account for the synergies identified between these two systems.

Portfolio HAM-3

Emergency Department Information System, Pharmacy Management, Dictation
Facilities with this portfolio: 1211

Emergency Department Information Systems (EDIS) are designed to automate and streamline the department's workflow and deliver patients a more efficient and improved quality of care. These systems are specifically designed to meet the unique needs of emergency room patients and physicians. Pharmacy management systems primarily manage data with respect to the dispensing of prescriptions. However, they also control inventory, assist with the billing of claims, and ensure compliance with laws and regulations. A hospital's dictation system allows physicians to create voice recordings. These recordings allow hands-free documentation of procedures in real-time as well as providing out-of-station physicians with the ability to leave patient instructions and orders. In the busy emergency room, all three of these systems may play a pivotal role in increasing the speed of care delivered to the patient which, in turn would impact quality outcomes.

Heart attack readmission	
	Estimate
OR_Scheduling	−0.767901
Lab_Outreach Services	−11.043843
Dictation with Speech Recognition	−5.556865
OR_Scheduling:Lab_Outreach	15.285874
OR_Scheduling:Dictation_SR	8.091687
OR_Scheduling:Lab_Outreach:Dictation_SR	3.208897

Portfolio HAR-1

Operating Room Scheduling, Laboratory Outreach Services
Facilities with this portfolio: 265

An operating room scheduling system provides physicians and administrators with information on each surgical procedure that is planned, currently underway, or has been completed. The system also assists with material management, material requirement planning, and pre-admission consultations. It also offers an opportunity

to record clinical notes for procedures, sterilization management, and transcription. Laboratory outreach is service offered by facilities where a facility's laboratory services are made available to outpatients as well as patients of other facilities and physicians. The laboratory outreach service can benefit the hospital by increasing revenues and filling unused capacity, as well as building relationships with patients and physicians in the community. However, while financial benefits are clear, laboratory outreach has not been traditionally associated with improved quality of care. Identifying a possible synergy is difficult between these two seemingly distantly-related areas of care. There may be a tie between the patients who receive laboratory services, and the results of those tests necessitating operative services.

Portfolio HAR-2

Operating Room Scheduling, Dictation with Speech Recognition
Facilities with this portfolio: 458

As mentioned earlier, a dictation system allows physicians to record patient instructions. However, an enhanced dictation system equipped with speech recognition allows the recordings to be converted into a digital format and easily imported into the computer as text. Conversely, the speech functions also allows patient statistics such date of birth, medical history and patient instructions to be transferred from the computer onto the recording. This system, together with the operating room scheduling system may offer a synergistic effect through automating note taking during procedures and thereby reducing documentation error.

Portfolio HAR-3

Operating Room Scheduling, Laboratory Outreach Services, Dictation with Speech Recognition
Facilities with this portfolio: 130

All three member systems have been introduced in previous portfolios. However, this portfolio has been adopted by relatively few facilities. In this combination of systems, we have effectively a merger of the previous two portfolios. The synergistic effects would be expected as laboratory outreach services and dictation with speech recognition have both demonstrated a positive synergy when associated with an operating room scheduling system.

Heart failure mortality	
	Estimate
Order_Entry	0.026469
Chart_Track	−0.703990
Laboratory_IS	0.185730
Microbiology	−1.446710
Order_Entry:Chart_Tracking	1.591157
Laboratory_IS:Microbiology	1.661938
	Estimate
Cardiology_Cath.Lab	−1.016270
Pharmacy_Management	−0.152552

Heart failure mortality	
	Estimate
Chart_Tracking	0.160954
Anatomical_Pathology	0.151822
Cardiology_Cath.Lab:Pharmacy_Management:	
Chart_Tracking:Anatomical_Pathology	1.592765

Portfolio HFM-1

Order Entry, Chart Tracking
Facilities with this portfolio: 1512

An order entry system is a component of an electronic medical records system which allows patient orders to be entered directly into the electronic record at the point of service. It also provides a mechanism to communicate those orders to external parties such as pharmacies and laboratories. These orders are typically communicated via encrypted Internet connections. A chart tracking system is also usually a module of a larger electronic medical records (EMR) management tool that is designed to manage the patient's paper-based records. Chart tracking systems can significantly streamline the processes and reduce the workload associated with records management. As sub-components of a common EMR, these systems are closely related, and therefore lend themselves to leveraging the other's benefits.

Portfolio HFM-2

Laboratory Information Systems, Microbiology
Facilities with this portfolio: 1514

Care givers use laboratory information systems (LIS) to manage an assortment of inpatient and outpatient medical testing, including hematology, chemistry, oncology, immunology and microbiology. As a specialized LIS, the microbiology system is designed to seamlessly integrate into the microbiology testing workflow enabling laboratories to achieve standardized, precise, and consistent results while maximizing lab efficiency. These two systems are also very closely related, and were present in a large percentage of hospitals analyzed. The overlap in functionality between these two systems may offer insight into intersystem synergies.

Portfolio HFM-3

Cardiology: Catheterization Laboratory, Pharmacy Management, Chart Tracking, Anatomical Pathology
Facilities with this portfolio: 463

The Cardiology: Catheterization Laboratory system collects, stores, maintains and protects still images and video created during cardiac catheterization procedures. These visual elements are necessary to maximize the efficacy of these procedures. Whereas the links between the pharmacy management and chart tracking system are not readily apparent, any synergy they provide the cardiology system would naturally enhance the heart failure mortality quality metric.

Heart failure readmission	
	Estimate
Blood_Bank	0.123467
Microbiology	−0.025624
Obstetrical_Systems	−0.558120
Radiology_MRI	−0.143150
Blood_Bank:Microbiology: Obstetrical_Systems:Radiology_MRI	2.672228
	Estimate
Blood_Bank	0.048369
Microbiology	0.127110
Consumer_Portal	1.142723
Anatomical_Pathology	1.627532
CPOE	0.373514
Blood_Bank:Consumer_Portal:CPOE	3.713268
Microbiology:Anatomical_Pathology:CPOE	4.281513

Portfolio HFR-1

Blood Bank, Microbiology, Obstetrical Systems, Radiology: Medical Resonance
    Imaging
Facilities with this portfolio: 820

A blood bank information system is a multi-module application that assists in areas
such as donor recruitment, blood collection, inventory control, donor testing, ship-
ping, transfusion, and billing. An obstetrical information system receives analog
information from various monitors which is digitized before being input. OB systems
typically have admission, transfer, edit, and discharge functions. This portfolio is
rather unique for two reasons. First, it is one of only two portfolios that incorporates
four IT systems; and second, these systems represent four distinct departments within
the healthcare facility. This portfolio offers further opportunity to explore the causes
behind the synergies observed.

Portfolio HFR-2

Blood Bank, Consumer Portal, Computerized Physician Order Entry
Facilities with this portfolio: 120

A consumer portal provides patients with direct access to their personal information
regarding their health plan coverage, medical history and treatment plans, as well
as offering patient services such as appointment scheduling and prescription refill
ordering. A computerized physician order entry system (CPOE) allows entering of
medication orders or other physician instructions electronically instead of on paper charts.
The use of a CPOE system can help reduce errors related to illegible handwriting or
transcription of medication orders. This recommended portfolio was found in the
fewest facilities in this study. It is also the only one to include the consumer portal. It is
possible that relative scarcity of consumer portals reflects untapped synergies.

Portfolio HFR-3

Microbiology, Anatomical Pathology, Computerized Physician Order Entry
Facilities with this portfolio: 336

An anatomical pathology laboratory information system (APLIS) logs specimens, records microscopic findings, regulates laboratory workflow, formulates reports, distributes them to the intended recipients throughout the healthcare system, and supports quality assurance measures. They also support asset tracking and digital imaging. The results entered into the microbiology system and APLIS are accessible to the physicians through their CPOE. Therefore promoting the accuracy of, and expediting the access to these test results may generate the synergies identified.

Pneumonia 30 day mortality	
	Estimate
Dictation	−0.314298
Abstracting	−0.257044
Radiology_Angiography	1.931669
Dictation:Abstracting	0.886240
	Estimate
Radiology_DM	−0.040562
Radiology_Nuclear	0.140742
Operating_Room_Pre	0.770414
OR_Scheduling	0.055803
Radiology_DM:Operating_Room_Pre:OR_Scheduling	3.855601

Portfolio PM-1

Dictation, Abstracting
Facilities with this portfolio: 1570

A coding and abstracting information system efficiently summarizes clinical data. The abstracting process supports later activities such as coding and reimbursement, quality improvement initiatives, billing audits, and clinical research. This two member-system portfolio was the most commonly found at the facilities under review. Interestingly it is also the only portfolio to include an abstracting information system. Both of these systems help expedite the administration function, and allow care providers to devote more of their time and attention to providing care to the patient.

Portfolio PM-2

Radiology: Digital Mammography, Operating Room: Pre-Operative, Operating Room Scheduling
Facilities with this portfolio: 660

A digital mammography system collects, stores, manages and disseminates x-ray images created during breast exams. The resulting images are analyzed for abnormalities which may indicate cancerous tissue. The pre-operative system assists anesthesiologists in pre-operative patient assessment and application of anesthesia. Pneumonia is of grave concern for patients recovering from surgery - particularly in the elderly.

Maximizing the quality of care during all stages of the surgical process may reduce the occurrence of pneumonia and thereby positively affect this metric.

Pneumonia readmission	
	Estimate
Radiology_DR	−0.883888
Dictation	−0.646310
Radiology_DF	1.330881
Operating_Room_Post	−1.059171
Radiology_DR:Dictation	0.375184
Radiology_DR:Radiology_DF:Operating_Room_Post	2.447464

Portfolio PR-1

Radiology: Digital Radiography, Dictation
Facilities with this portfolio: 1297

A digital radiography system offers advancement over the traditional film x-ray. Images are held digitally and are available immediately. This eliminates the need to wait for film development, and allows physicians to more quickly review and diagnose patients. Since a chest x-ray is the primary means by which physicians diagnose pneumonia, it stands to reason that any tool which improves upon the functionality or speed of this treatment would positively affect a facilities performance in the frequency of patients readmitted because of pneumonia.

Portfolio PR-2

Radiology: Digital Radiography, Radiology: Digital Fluoroscopy, Operating Room: Post-Operative
Facilities with this portfolio: 940

Digital fluoroscopy is a digital x-ray imaging system similar to digital radiography; however, the images are dynamic. Digital fluoroscopy is a form of x-ray that allows physicians to inspect deep tissues in the body in real-time on a computer monitor. It provides detailed images of the function and structure of areas like the lungs, the liver, the heart and kidneys. A post-operative care system can give consultative and decision support to surgical recovery staff with the goal of reducing surgical site infections, heart attacks, blood clots, and postoperative pneumonia. Once again, this portfolio contains systems which directly relate to either the diagnosis or prevention of pneumonia. Therefore, the synergistic effects identified are not surprising.

9.4.2 Negative Synergies Identified

Interestingly, synergistic effects can affect patient quality outcomes in both a positive as well as negative manner. While most portfolio coefficients revealed little to no synergistic effects, and several portfolios (as detailed above) indicated a positive synergistic effect, four portfolios reported a negative impact on quality. Three of these

portfolios apply to the heart attack readmission metric, and the fourth deals with heart failure readmission.

Heart attack readmission	
	Estimate
Operating_Room_Scheduling	3.044564
Radiology_DR	2.808544
EDI	2.752098
CPOE	1.803123
Operating_Room_Scheduling:EDI	−3.535674
Operating_Room_Scheduling:CPOE	−4.989675
Radiology_DR:CPOE	−8.012095

In these results we see that the operating room scheduling, radiology: digital radiography, electronic data interchange (EDI), and computerized physician order entry (CPOE) have been combined into a single portfolio. An EDI system allows the transfer of information between two disparate systems of networks (2013). These tools are often used to allow legacy within a facility to communicate or allow the transfer of patient records between facilities. Each of these systems have a moderately positive correlation with better than average results. However, when operating room scheduling is joined with EDI or CPOE the combined scores are significantly negative. Likewise, when radiology: digital radiography is joined with CPOE, we see an even greater change to the results. Since most of these systems appear in one or more of the candidate portfolios, we cannot simply dismiss the systems as offering little value in a portfolio environment. However, it is clear that for at least the heart attack readmission quality metric, intersystem dynamics are present which may be hampering quality.

Heart failure readmission	
	Estimate
Blood_Bank	0.048369
Microbiology	0.127110
Consumer_Portal	1.142723
CPOE	0.373514
Blood_Bank:Microbiology:Consumer_Portal:CPOE	−4.927736

The heart failure readmission results above reveal that the blood bank, microbiology, consumer portal and CPOE systems have nearly a neutral influence on quality. However, when all four systems are combined, they offer a strong negative impact. The commonality between these two examples is the inclusion of CPOE in the portfolios. As documented by Koppel et al. (2005), CPOE systems produce the opportunity to introduce medical errors into the system, and thus negatively impact quality.

The causes of these negative effects are not fully apparent. However, what is clear from these results is that the introduction of additional IT systems into a healthcare environment may not always prove to be advantageous, and in some cases may result in a detriment to patients and the organization. This finding directly supports the Yu and Houston (2007) contention discussed earlier that IT adoption is not a strong predictor of quality performance.

9.5 Discussion

The results outlined previously illustrate that synergistic effects are occurring between multiple IT systems within the healthcare arena. The purpose of this research was to introduce an approach to identify portfolios that harness these synergies and to provide a mechanism to confirm their existence. Specifically addressing the research objectives:

Research Objective 1
Identify optimal portfolios of information technology that are positively associated with above average quality performance at healthcare organizations.

As identified earlier, the data mining phase of the approach identified many portfolios associated with *better than national average* quality results. Many of these portfolios contained systems which were individually associated with *above average results*. It would be expected that when combining these systems into a portfolio, the resulting accumulative effect on quality would also be positive. Therefore, a portfolio's positive correlation with better than the national average is not sufficient to predict that it contains a synergistic effect. However, these portfolios would be suitable candidates for further analysis using logistic regression.

Research Objective 2
Identify if synergistic effects exist between the components of the optimal IT portfolio. Specifically, are individual technologies more positively associated with quality when used in conjunction with other technologies within an IT portfolio than when used in isolation?

Using the second step of the recommended approach, logistic regression we were able to support the presence of synergistic effects between select IT systems. These synergistic effects were specific to individual quality metrics and their effects seem to be mitigated by the presence, or lack of presence, of other IT systems.

9.6 Research Contribution and Impact

The results of this research have significant implications for both theory and practice. The exploration of optimal portfolios and synergistic effects adds to the knowledge base on impact of portfolios on organizational performance by extending it to the case of healthcare and healthcare quality. By applying the portfolio theory to information technology investments within the healthcare context, insights have been gained into a lightly explored subject area using concepts rarely applied in this arena.

Specifically, contributions from this research include:

1. A clearer understanding of HIT's impact on quality, and therefore may help guide decision-makers when planning and implementing future IT investments. Healthcare administrators seeking to bolster or maximize a particular patient outcome for their

facility, can compare their current IT system mix to those recommended portfolios, and identify those systems which may provide the greatest return on investment.

2. Understanding the inter-system synergies will guide strategic planners of facilities based on systems previously adopted. Those facilities with the recommended systems already in place, but that are not performing well on the quality metric, will have additional information to inform their performance improvement efforts.

3. The identification of systems that have no, or relatively minor, impact on quality may inform the design of future versions of these systems. Identifying combinations of systems with lower than expected interactions can aid HIT system vendors seeking to enhance their offerings by providing an area of focus for future development.

4. A unique application of Portfolio Theory. To date, the Portfolio Theory has been used extensively but almost exclusively within the finance arena (Bridges et al. 2002). Extending the application of this well-defined and well-understood theory to the healthcare domain supports the validity of this research while also expanding the usefulness of the theory.

5. Interdisciplinary Approach: bridging three domains. The approach outlined by this research draws from three independent domains. The Portfolio Theory is borrowed from the economics and finance, the data mining techniques are drawn from information technology, and the examination of HIT systems reflects the healthcare domain.

9.7   Research Limitations

The data used for this research was provided to the public in the form of two datasets. Neither dataset was designed specifically for this research. Therefore, the data structure and granularity were not ideal. The process of ranking the facilities by their quality metric into the three classifications (*above the national average, equal to the national average or below the national average*) was not fully detailed. The dataset documentation did not indicate through what mechanisms these facilities were assigned their rating. Furthermore, a large majority of the facilities obtained an *equal to the national average rating* on each metric. This indicates the parameters for this rating must have been rather large. If facilities were ranked into more than three categories, the requirements to receive an average ranking were constrained, or if the facilities performance was reported as a numerical value, greater precision could be attained.

The IT systems reported in the HIMSS dataset did not include extent of system use or user training levels which would helpful to combat endogeneity concerns. However, we believe the large sample size still provides realistic averages.

More current, comprehensive, accurate and robust datasets with finer granularity are available, and will continue to be made available. As future generations of systems advance and mature, their synergistic relationships will surely evolve. Therefore regular application of this approach to updated datasets will be required.

Biography

Ken Pinaire   Ken has been teaching at the graduate level since 2002. His areas of interest include promoting healthcare quality through the use of technology and data mining. He seeks to assist healthcare systems maximize the value of their technology through targeted investments. He is currently studying the synergistic effects of information technology systems in the healthcare arena.

Surendra Sarnikar   is an Associate Professor in Information Systems at the College of Business and Information Systems Dakota State University. He holds a Bachelors degree in Engineering from Osmania University India and a PhD in Management Information Systems from the University of Arizona. He teaches healthcare informatics design research and knowledge management at the Dakota State University. He has published several conference and Journal publications in the area of healthcare information systems knowledge management systems and information retrieval.

References

Adams, W. G., Mann, A. M., & Bauchner, H. (2003). Use of an electronic medical record improves the quality of urban pediatric primary care. *Pediatrics, 111*(3), 626–632.

Ancker, J. S., Kern, L. M., Abramson, E., & Kaushal, R. (2012). The triangle model for evaluating the effect of health information technology on healthcare quality and safety. *Journal of the American Medical Informatics Association*, *19*(1), 61–65.

Bardach, N. S., Huang, J., Brand, R., & Hsu, J. (2009). Evolving health information technology and the timely availability of visit diagnoses from ambulatory visits: A natural experiment in an integrated delivery system. *BMC Medical Informatics and Decision Making, 9*, 35.

Bendoly, E., Rosenzweig, E. D., & Stratman, J. K. (2009). The efficient use of enterprise information for strategic advantage: A data envelopment analysis. *Journal of Operations Management, 27*(4), 310–323.

Bhattacherjee, A., Hikmet, N., Menachemi, N., Kayhan, V. O., & Brooks, R. G. (2007). The differential performance effects of healthcare information technology adoption. *Information Systems Management, 24*(1), 5–14.

Bridges, J. F. P., Stewart, M., King, M. T., & van Gool, K. (2002). Adapting portfolio theory for the evaluation of multiple investments in health with a multiplicative extension for treatment synergies. *The European Journal of Health Economics, 3*(1), 47–53.

Burke, D. E., & Menachemi, N. (2004). Opening the black box: Measuring hospital information technology capability. *Health Care Management Review, 29*(3), 207–217.

Burke, D. E., Wang, B. B. L., Wan, T. T. H., & Diana, M. L. (2002). Exploring hospitals' adoption of information technology. *Journal of Medical Systems, 26*(4), 349–355.

Campbell, E., Li, H., Mori, T., Osterweil, P., & Guise, J. M. (2008). The impact of health information technology on work process and patient care in labor and delivery. In K. Henriksen, J. B. Battles, M. A. Keyes, et al. (Eds.), Advances in patient safety: New directions and alternative approaches (Vol. 4: Technology and medication safety). Rockville: Agency for Healthcare Research and Quality (US).

Campion, T., Edwards, A., Johnson, S., & Kaushal, R. (2013). Health information exchange system usage patterns in three communities: Practice sites, users, patients, and data. *International Journal of Medical Informatics, 82*(9), 810–820.

Chatterjee, S., Chakraborty, S., Sarker, S., Sarker, S., & Lau, F. Y. (2009). Examining the success factors for mobile work in healthcare: A deductive study. *Decision Support Systems, 46*(3), 620–633.

Cresswell, K. M., Worth, A., & Sheikh, A. (2010). Actor-network theory and its role in understanding the implementation of information technology developments in healthcare. *BMC Medical Informatics and Decision Making, 10*, 67.

Cochran, M. B., Snyder, R. R., Thomas, E., Freeman, D. H., & Hankins, G. D. (2011). Implementation of health information technology to maximize efficiency of resource utilization in a geographically dispersed prenatal care delivery system. *American Journal of Perinatology, 29*(4), 251–258.

Cohen, A. N., Chinman, M. J., Hamilton, A. B., Whelan, F., & Young, A. S. (2013). Using patient-facing kiosks to support quality improvement at mental health clinics. *Medical Care, 51*(3 Suppl 1), 13–20.

Davis, M. A., & Pavur, R. J. (2011). The relationship between office system tools and evidence-based care in primary care physician practice. *Health Services Management Research, 24*(3), 107–113.

Deily, M. E., Hu, T., Terrizzi, S., Chou, S. Y., & Meyerhoefer, C. D. (2013). The impact of health information technology adoption by outpatient facilities on pregnancy outcomes. Health Services Research, 48(1), 70–94.

DesRoches, C. M., Campbell, E. G., Vogeli, C., Zheng, J., Rao, S. R., Shields, A. E., Donelan, K., Rosenbaum, S., Bristol, S. J., & Jha, A. K. (2010). Electronic health records' limited successes suggest more targeted uses. *Health Affairs (Millwood), 29*(4), 639–646.

Dorenfest, S. (2000). The decade of the '90s. Poor use of it investments contributes to growing healthcare crisis. *Healthcare Informatics, 17*(8), 64–67.

El-Kareh, R., Gandhi, T. K., Poon, E. G., Newmark, L. P., Ungar, J., Lipsitz, S. D., et al. (2009). Trends in primary care clinician perceptions of a new electronic health record. *Journal of General Internal Medicine, 24*(4), 464–468.

Flom, P., & Strauss, S. M. (2003). Some graphical methods for interpreting interactions in logistic and Ols regression. *Multiple Linear Regression Viewpoints, 29*(1), 1–7.

Fraenkel, D. J., Cowie, M., & Daley, P. (2003). Quality benefits of an intensive care clinical information system. *Critical Care Medicine, 31*(1), 120–125.

Friesner, D. L., Rosenman, R., & McPherson, M. Q. (2007). Does a single case mix index fit all hospitals? Empirical evidence from Washington state. *Research in Healthcare Financial Management, 11*(1), 35–55.

Frimpong, J. A., Jackson, B. E., Stewart, L. M., Singh, K. P., Rivers, P. A., & Bae, S. (2013). Health information technology capacity at federally qualified health centers: A mechanism for improving quality of care. *BMC Health Services Research, 13*(1), 35.

Furukawa, M., & Adam, T. (2008). Health information technology and hospital quality of care. AMIA Annual Symposium Proceedings, 864.

Gaylin, D. S., Moiduddin, A., Mohamoud, S., Lundeen, K., & Kelly, J. A. (2011). Public attitudes about health information technology, and its relationship to health care quality, costs, and privacy. *Health Services Research, 46*(3), 920–938.

Glossary of Terms and Acronyms Related to e-Health. (2013), from. http://www.health.state.mn.us/e-health/glossary.html

Gluck, M. E. (2009). Is health information technology associated with patient safety in the United States? *Findings Brief, 12*(3), 1–3.

Golob, J. F., Jr., Fadlalla, A. M., Kan, J. A., Patel, N. P., Yowler, C. J., & Claridge, J. A. (2008). Validation of surgical intensive care-infection registry: A medical informatics system for intensive care unit research, quality of care improvement, and daily patient care. *Journal of the American College of Surgeons, 207*(2), 164–173.

Harrell, F. E. (2001). *Regression modeling strategies*. New York: Spinger.

Hazlehurst, B., McBurnie, M. A., Mularski, R. A., Puro, J. E., & Chauvie, S. L. (2012). Automating care quality measurement with health information technology. *The American Journal of Managed Care, 18*(6), 313–319.

Hennington, A. H., & Janz, B. D. (2007). Information systems and helathcare XVI: Physician adoption of electronic meidcal records: Applying the UTAUT model in a healthcare context. *Communications of AIS, 2007*(19), 60–80.

Hooper, D. K., Kirby, C. L., Margolis, P. A., & Goebel, J. (2013). Reliable individualized monitoring improves cholesterol control in kidney transplant recipients. *Pediatrics, 131*(4), e1271–e1279.

Hunt, J. S., Siemienczuk, J., Gillanders, W., LeBlanc, B. H., Rozenfeld, Y., Bonin, K., & Pape, G. (2009). The impact of a physician-directed health information technology system on diabetes outcomes in primary care: A pre- and post-implementation study. *Informatics in Primary Care, 17*(3), 165–174.

Jamal, A., McKenzie, K., & Clark, M. (2009). The impact of health information technology on the quality of medical and health care: A systematic review. *Health Information Management Journal, 38*(3), 26–37.

Jean-Jacques, M., Persell, S. D., Thompson, J. A., Hasnain-Wynia, R., & Baker, D. W. (2011). Changes in disparities following the implementation of a health information technology-supported quality improvement initiative. *Journal of General Internal Medicine, 27*(1), 71–77.

Kern, L. M., Dhopeshwarkar, R., Barron, Y., Wilcox, A., Pincus, H., & Kaushal, R. (2009). Measuring the effects of health information technology on quality of care: A novel set of proposed metrics for electronic quality reporting. *Joint Commission Journal on Quality and Patient Safety, 35*(7), 359–369.

Koppel, R., Metlay, J. P., Cohen, A., Abaluck, B., Localio, A. R., Kimmel, S. E., & Strom, B. L. (2005). Role of computerized physician order entry systems in facilitating medication errors. *Jama-Journal of the American Medical Association, 293*(10), 1197–1203.

Lin, B.-W., Chen, C.-J., & Wu, H.-L. (2006). Patent portfolio diversity, technology strategy, and firm value. *IEEE Transactions on Engineering Management, 53*(1), 17–26.

Loiselle, C. G., Edgar, L., Batist, G., Lu, J., & Lauzier, S. (2010). The impact of a multimedia informational intervention on psychosocial adjustment among individuals with newly diagnosed breast or prostate cancer: A feasibility study. *Patient Education and Counseling, 80*(1), 48–55.

Lucero, R. J., Ji, H., de Cordova, P. B., & Stone, P. (2011). Information technology, nurse staffing, and patient needs. *Nursing Economics, 29*(4), 189–194.

McCullough, J. S., Casey, M., Moscovice, I., & Prasad, S. (2010). The effect of health information technology on quality in U.S. hospitals. *Health Aff (Millwood), 29*(4), 647–654.

Menachemi, N., Burkhardt, J., Shewchuk, R., Burke, D., & Brooks, R. G. (2006). Hospital information technology and positive financial performance: A different approach to finding an Roi. *Journal of Healthcare Management/American College of Healthcare Executives, 51*(1), 40–58.

Menachemi, N., Chukmaitov, A., Saunders, C., & Brooks, R. G. (2008). Hospital quality of care: Does information technology matter? The relationship between information technology adoption and quality of care. *Health Care Management Review, 33*(1), 51–59.

Menon, N. M., Lee, B., & Eldenburg, L. (2000). Productivity of information systems in the healthcare industry. *Information Systems Research, 11*(1), 83.

Monegain, B. (2009). Global market for hospital it systems pegged at $35b by 2015. *Healthcare IT News*.

Morin, P. C., Wolff, L. T., Eimicke, J. P., Teresi, J. A., Shea, S., & Weinstock, R. S. (2009). Record media used by primary care providers in medically underserved regions of upstate New York was not pivotal to clinical result in the Informatics for Diabetes Education and Telemedicine (IDEATel) project. *Informatics in Primary Care, 17*(2), 103–112.

Myung Ko, K. A., & Osei-Bryson, K.-M. (2004). Using regression splines to assess the impact of information technology investments on productivity in the health care industry. *Information Systems Journal, 14*(1), 43–63.

Nirel, N., Rosen, B., Sharon, A., Samuel, H., & Cohen, A. D. (2011). The impact of an integrated hospital-community medical information system on quality of care and medical service utilisation in primary-care clinics. *Informatics for Health and Social Care, 36*(2), 63–74.

Nirel, N., Rosen, B., Sharon, A., Blondheim, O., Sherf, M., Samuel, H., et al. (2009). The impact of an integrated hospital-community medical information system on quality and service utilization in hospital departments. *International Journal of Medical Informatics, 79*(9), 649–657.

O'Connor, P. J., Crain, A. L., Rush, W. A., Sperl-Hillen, J. M., Gutenkauf, J. J., & Duncan, J. E. (2005). Impact of an electronic medical record on diabetes quality of care. *Annals of Family Medicine, 3*(4), 300–306.

Pare, G., & Sicotte, C. (2001). Information technology sophistication in health care: An instrument validation study among Canadian hospitals. *International Journal of Medical Informatics, 63*(3), 205–223.

Piontek, F., Kohli, R., Conlon, P., Ellis, J. J., Jablonski, J., & Kini, N. (2010). Effects of an adverse-drug-event alert system on cost and quality outcomes in community hospitals. *American Journal of Health-System Pharmacy, 67*(8), 613–620.

Pizzi, R. (2007). U.S. Ehr market to approach $5 billion by 2015. *Healthcare IT News*.

Restuccia, J. D., Cohen, A. B., Horwitt, J. N., & Shwartz, M. (2012). Hospital implementation of health information technology and quality of care: Are they related? *BMC Medical Informatics and Decision Making, 12*, 109.

Roberts, L. L., Ward, M. M., Brokel, J. M., Wakefield, D. S., Crandall, D. K., & Conlon, P. (2010). Impact of health information technology on detection of potential adverse drug events at the ordering stage. *American Journal of Health-System Pharmacy, 67*(21), 1838–1846.

Romano, M. J., & Stafford, R. S. (2011). Electronic health records and clinical decision support systems: Impact on national ambulatory care quality. *Archives Internal Medicine, 171*, 897–903.

Savage, I., Cornford, T., Klecun, E., Barber, N., Clifford, S., & Franklin, B. D. (2010). Medication errors with electronic prescribing (Ep): Two views of the same picture. *BMC Health Services Research, 10*, 135–142.

Setia, P., Setia, M., Krishnan, R., & Sambamurthy, V. (2011). The effects of the assimilation and use of it applications on financial performance in healthcare organizations. *Journal of the Association for Information Systems, 12*(3), 274–298.

Sharkey, S., Hudak, S., Horn, S. D., Barrett, R., Spector, W., & Limcangco, R. (2013). Exploratory study of nursing home factors associated with successful implementation of clinical decision support tools for pressure ulcer prevention. *Advances in Skin & Wound Care, 26*(2), 83–92.

Shelley, D., Tseng, T. Y., Matthews, A. G., Wu, D., Ferrari, P., Cohen, A., et al. (2011). Technology-driven intervention to improve hypertension outcomes in community health centers. *The American Journal of Managed Care, 17*(12 Spec No.), 103–110.

Spielberg, F., Kurth, A., Reidy, W., McKnight, T., Dikobe, W., & Wilson, C. (2011). Iterative evaluation in a mobile counseling and testing program to reach people of color at risk for HIV–new strategies improve program acceptability, effectiveness, and evaluation capabilities. *AIDS Education and Prevention, 23*(Suppl 3), 110–116.

Swanson Kazley, A., & Diana, M. L. (2011). Hospital computerized provider order entry adoption and quality: An examination of the United States. *Health Care Management Review, 36*(1), 86–94.

Thrasher, E. H., Byrd, T. A., & Hall, D. (2006). Information systems and healthcare XV: Strategic fit in healthcare integrated delivery systems: An empirical investigation. *Communications of AIS, 2006*(18), 692–709.

Thrasher, E. H., Craighead, C. W., & Byrd, T. A. (2010). An empirical investigation of integration in healthcare alliance networks. *Decision Support Systems, 50*(1), 116–127.

Valdmanis, V. G., Rosko, M. D., & Mutter, R. L. (2008). Hospital quality, efficiency, and input slack differentials. *Health Services Research, 43*(5 part 2), 1830–1848.

Virga, P. H., Jin, B., Thomas, J., & Virodov, S. (2012). Electronic health information technology as a tool for improving quality of care and health outcomes for HIV/AIDS patients. *International Journal of Medical Informatics, 81*(10), 39–45.

Wakefield, D. S., Ward, M. M., Loes, J. L., O'Brien, J., & Sperry, L. (2010). Implementation of a telepharmacy service to provide round-the-clock medication order review by pharmacists. *American Journal of Health-System Pharmacy, 67*(23), 2052–2057.

Yu, F., & Houston, T. K. (2007). Do "most wired" hospitals deliver better care? *The Joint Commission Journal on Quality and Patient Safety, 33*(3), 136–144.

Zhu, K. (2004). The complementarity of information technology infrastructure and e-commerce capability: A resource-based assessment of their business value. *Journal of Management Information Systems, 21*(1), 167–202.

Chapter 10
Introduction: Research-in-Progress Studies

Thilini Ariyachandra and Amit V. Deokar

Abstract Keeping with the "Reshaping Society" theme of the ICIS 2013 conference, the Pre-ICIS SIGDSS workshop sought forward-thinking research in the areas of analytics, collaboration and decision support with special focus on business intelligence and social media. The track aimed to promote theoretical, design science, behavioral research and emerging applications in innovative areas of analytics, collaboration and decision support. The Research-in-Progress work from the workshop, points to the potential of BI, DSS and Analytics technologies to influence quality of life and business such as improve consumer purchase decision making or patient decision making in healthcare. This work has been summarized in this chapter.

Keywords Research • Analytics • Social media • Impact • Society

Business Intelligence (BI)/ Decision Support (DSS)/ Analytics have become core to many businesses. The social media scape has come to modify and redefine not only businesses but also societal behavior and practices. Although addressed by research in the past few years, these domains are still evolving. Traditional approaches to collaboration and decision support are experiencing evolution and growth from BI and social media technologies and trends. As organizations in various sectors formulate IT strategies and investments, it is imperative to understand how various emerging technologies under the BI/DSS umbrella such as big data, mobile and wireless technologies, cloud computing, and recent collaboration tools can be used effectively for decision making in these organizations. For instance, the explosive growth in big data and social media analytics requires examination of the impact of these technologies and applications on decision making and society.

Research Track Coordinators: Thilini Ariyachandra, Amit Deokar

T. Ariyachandra (✉)
Xavier University, 3800 Victory Parkway, Cincinnati, OH 45207, USA
e-mail: ariyachandrat@xavier.edu

A.V. Deokar
Pennsylvania State University, 5101 Jordan Rd., Burke Center 268, Erie, PA 16563, USA
e-mail: amit.deokar@psu.edu

The factors examined during decision making have considerably expanded and the window for decision making has reduced due to the immense growth in data in varying structure, complexity, velocity and quantity and the advances in analytics. Analyzing, organizing, understanding the patterns and trends that lead to prediction from big data has become critical to organizations. Beyond prediction, big data analytics has created the potential to change the future and reshape society by providing data driven findings that could improve healthcare choices, policy and governance and influence economies. The full potential of prescriptive analytics to influence and shape business and society is yet to be discovered. The convergence of new analytic technologies and big data has enabled innovative organizations to move from data reporting and diagnostics to prediction and prescription.

Keeping with the "Reshaping Society" theme of the ICIS conference, the Pre-ICIS SIGDSS workshop sought forward-thinking research in the areas of analytics, collaboration and decision support with special focus on business intelligence and social media. The track aimed to promote theoretical, design science, behavioral research and emerging applications in innovative areas of analytics, collaboration and decision support. Three main themes emerged in the research in progress papers presented at the track. They are BI and organizational competitiveness, BI and healthcare, and BI and commerce. The research that addresses BI and organizational competitiveness is described first.

Competing in an environment with a constant influx of data requires agility. Infusing agility into an enterprise level business intelligence infrastructure is challenging. Knabke and Olbrich (2013) examine the characteristics of BI that can impact its agility and organizational competitiveness. Specifically they investigate if technological trends and methods that facilitate agility such as in-memory databases and extreme programming can help to achieve more agility in BI. Creating and sharing knowledge among members via organizational social networks is recognized as another means of gaining sustained competitive advantage. Using an agent based simulation model, Song and Choi (2013) attempt to provide a deeper understanding of the dynamics of organizational network structures within its task environments that can enable innovative organizational performance. Finally, through an extensive literature review, Cattaneo et al. (2013) presents a model for assessing technology intelligence capabilities that enables organizations to review the capabilities a company needs to manage discontinuous technological change.

Under the BI and healthcare theme, the importance and use of clinical decision support systems (CDSS) to assist in clinical decisions and productivity are investigated by Han et al. (2013). Specifically, they investigate the impact of practice size and meaningful use engagement on CDSS usage, and the effect of CDSS usage on productivity over time. Khanal et al. (2013), describes the use of an interactive collaborative virtual team training system that integrates multisensory devices for use in Advanced Cardiac Life Support to enable remote, ubiquitous team training. Patient decision aids that are individualized to reflect personal information needs and decision making desires that improve healthcare system effectiveness is the focus of Mortony and Sarnikar (2013). They propose an

intelligent user experience formula to help establish a template for future individualization of patient decision aids.

Finally, within BI and commerce, an attentive in-store mobile recommender system, integrated into the user's glasses for use during purchase decisions, is described (Pfeiffer et al. 2013). The initial results from eye-tracking data during purchase decision making at a supermarket is discussed to help characterize and identify the different purchase decision contexts. Alic (2013) addresses the research gap in financial institution market surveillance that integrates both structured and unstructured user-generated content data with the information provided by the regulatory authority in a system. The author proposes a design theory to help formulate clear requirements for an IS system that supports market surveillance. Understanding how organizations can attract and retain user contributions for online collaborative problem solving efforts is investigated by Nguyen et al. (2013). They suggest the use of Flow Theory to provide useful guidance to design an engaging crowdsourcing experience.

The research in progress from the workshop, points to the potential of BI, DSS and Analytics technologies to influence quality of life and business such as improve consumer purchase decision making or patient decision making in healthcare. The research suggests the interesting developments yet to be identified and developed in the future in this field. Some of research in progress presented during the workshop is presented in this section.

Biography

Thilini Ariyachandra   is an Associate Professor of Management Information Systems in the Williams College of Business at Xavier University in Cincinnati Ohio USA. She received her Ph.D. from the Terry School of Business at the University of Georgia. Her research is focused on business intelligence Big Data and business analytics implementation and success. She has published in journals such as Decision Support Systems Communications of the AIS Communications of the ACM the Business Intelligence Journal and the International Journal of Business Intelligence Research. She serves on advisory boards of academic alliances such as the Teradata University Network and Microsoft Dynamics Academic Alliance. She also serves as the Program Chair for the Special Interest Group for Decision Support and Analytics for the Association for Information Systems.

Amit V. Deokar   is an Assistant Professor of Management Information Systems in the Sam and Irene Black School of Business at Pennsylvania State University Erie. His recent research interests are in decision support and analytics business process management and collaboration processes and technologies. His work has appeared in journals including Journal of Management Information Systems Communications of the AIS Information Systems Frontiers The DATA BASE for Advances in Information Systems and IEEE Transactions. He has also presented at national and

international conferences and authored book chapters in the field of Information Systems. He holds a BE in Mechanical Engineering from V. J. Technological Institute Mumbai a MS in Industrial Engineering from the University of Arizona and a PhD in Management Information Systems from the University of Arizona. He is currently serving as the Program Chair-elect of the AIS Special Interest Group on Decision Support and Analytics.

References

Alic, I. (2013). Market surveillance DSS: Towards an explanatory design theory. In *Proceedings of the Pre-ICIS SIGDSS Research Workshop*, Milan, Italy.

Cattaneo, G., Battistini, B., & Hacklin, F. (2013). Revisiting intelligence for technology and innovation management: An integrative review and assessment model. In *Proceedings of the Pre-ICIS SIGDSS Research Workshop*, Milan, Italy.

Han, W., Sharman, R., Sidiqqi, H., Singh, R., & Singh, G. (2013). The impact of practice size and meaningful use engagement on the usage of clinical decision support systems and practice productivity. In *Proceedings of the Pre-ICIS SIGDSS Research Workshop*, Milan, Italy.

Khanal, P., Parab, P., Gupta, A., Kahol, K., & Smith, M. (2013). Evaluating the performance of virtual worlds for collaborative time-critical medical training. In *Proceedings of the Pre-ICIS SIGDSS Research Workshop*, Milan, Italy.

Knabke, T., & Olbrich, S. (2013). Reconciliation of business intelligence principles and enterprise agility: First insights into an empirical study. In *Proceedings of the Pre-ICIS SIGDSS Research Workshop*, Milan, Italy.

Motorny, S., & Sarnikar, S. (2013). Design of an intelligent patient decision aid based on individual decision making styles and information need preferences. In *Proceedings of the Pre-ICIS SIGDSS Research Workshop*, Milan, Italy.

Nguyen, C., Onook Oh, O., Alothaim, A., de Vreede, T., & de Vreede, G. (2013). Engaging with online crowd: A flow theory approach. In *Proceedings of the Pre-ICIS SIGDSS Research Workshop*, Milan, Italy.

Pfeiffer, J., Pfeiffer, T., & Meißner, M. (2013). Towards attentive in-store recommender systems: Detecting explorative vs. goal-oriented decisions. In *Proceedings of the Pre-ICIS SIGDSS Research Workshop*, Milan, Italy.

Song, S., & Choi, S. (2013). Modeling dynamic organizational network structure. In *Proceedings of the Pre-ICIS SIGDSS Research Workshop*, Milan, Italy.

Chapter 11
Towards Attentive In-Store Recommender Systems

Jella Pfeiffer, Thies Pfeiffer, and Martin Meißner

Abstract We present research-in-progress on an attentive in-store mobile recommender system that is integrated into the user's glasses and worn during purchase decisions. The system makes use of the Attentive Mobile Interactive Cognitive Assistant (AMICA) platform prototype designed as a ubiquitous technology that supports people in their everyday-life. This paper gives a short overview of the technology and presents results from a pre-study in which we collected real-life eye-tracking data during decision processes in a supermarket. The data helps us to characterize and identify the different decision contexts based on differences in the observed attentional processes. AMICA provides eye-tracking data that can be used to classify decision-making behavior in real-time to make a recommendation process context-aware.

J. Pfeiffer (✉)
Institute of Information Systems and Marketing (IISM), Karlsruhe Institute of Technology (KIT) Englerstr. 14, 76131 Karlsruhe, Germany
e-mail: jella.pfeiffer@kit.edu

T. Pfeiffer
Center of Excellence Cognitive Interaction Technology, Bielefeld University,
P.O. Box 10 01 31, 33501 Bielefeld, Germany
e-mail: tpfeiffe@techfak.uni-bielefeld.de; http://www.techfak.de/~tpfeiffe/

M. Meißner
Department of Environmental and Business Economics, University of Southern Denmark,
Niels Bohrs Vej 9, DK-6700 Esbjerg, Denmark

Department of Marketing, Monash University, Melbourne, Australia
e-mail: meissner@sam.sdu.dk; http://findresearcher.sdu.dk:8080/portal/en/person/meissner

11.1   Introduction

The emergence of a new type of users called the digital natives has posed new questions in the field of information systems (IS). Digital natives have grown up surrounded by information and communication technology that is pervasive and ubiquitous (Prensky 2001; Tapscott 2008). As we are convinced that digital natives are the future users of decision support systems (DSS), we like to present a ubiquitous in-store recommender system.

Our work presented here is part of a larger project on a ubiquitous DSS (UDSS) called Attentive Mobile Interactive Cognitive Assistant (AMICA). The aim of that project is to realize AMICA as an attentive recommender system which is worn by the user during in-store purchase decisions. Such a system is context-aware, which means that eye-tracking technology is applied to learn from the users' attentional processes when standing in front of a product shelf. The system will automatically detect the user's needs and provide appropriate product information and recommendations.

Recently published articles have argued that a challenge for existing recommender systems is to elicit preference information in a minimally intrusive manner in order to reduce the user's effort (Ansari et al. 2000; Murray et al. 2010). A solution is to use UDSSs that are context-aware (Adomavicius et al. 2005; Adomavicius and Tuzhilin 2011; Lee and Benbasat 2010; Palmisano et al. 2008). Such systems learn the user's preferences in real-time and thus take into account that users construct their preferences while they process the decision-relevant information (Häubl and Murray 2003; Slovic 1995). Furthermore, previous research shows that not only users' preferences but also their decision strategies are highly contingent upon the context (Bettman et al. 1998; Payne et al. 1993). UDSS, such as AMICA, have the advantage that in real-world decision environments rich data can be gathered about the context, for example location information, eye movements, gestures and speech.

This paper gives a short overview of the technology to be introduced and presents results from a pre-study to convey the underlying concepts. It explains the AMICA design, which is based on specific user requirements for ubiquitous systems, and presents a first idea on how to automatically detect the decision context.

In the following sections, we provide a literature overview about recommender systems for in-store purchase decisions and suggest new concepts to make them context-aware. Then, we illustrate how our system architecture achieves specific requirements for making the DSS ubiquitous. Finally, we conducted a pre-study in which users evaluate a prototype of the system. This proof-of-concept demonstrates how the system can learn about users' information needs from the context.

11.2   In-Store Decision Support with Mobile Devices

Several publications have developed techniques for identifying products which are of interest for users. Those DSS display product information on mobile devices (Resatsch et al. 2008; van der Heijden 2006; von Reischach et al. 2009). The most

often used techniques are manually typing in a barcode or the product name, automatic barcode recognition, near field communication (NFC) and RFID. NFC — as the fastest method—achieves the highest perceived ease of use (von Reischach et al. 2009). All approaches require the user to get into close proximity of the target or even to pick it up and turn it around. NFC, for example, requires the user to hold the mobile device in a 5 cm range from the product and it requires on average 3.3 s for detecting the product. Considering that more than 100 different products of muesli are in a standard shelf in the supermarket, this highly manual interaction is tedious and unfeasible except when only information for a significantly small number of products is required. This, however, is contrasted with the problem that all described techniques are only collecting information about the products themselves but not about the topology of the product arrangement in the shelf. So no further orientation help for the user can be given and thus these systems cannot help during the visual search process itself. In summary, these approaches focus on displaying product information and/or provide simple recommendations made by other consumers or experts. The approaches are non-personalized, non-social and not very interactive. They require explicit input about the user's context and have very limited context-awareness overall.

Resatsch et al. (2008) found that digital natives valued their in-store decision support system more than digital immigrants. Moreover, they were interested in receiving recommendations from the system in addition to just getting product information. Furthermore, privacy and data security were not a concern for those respondents. However, the credibility of the information source was very important to them. Lee and Benbasat (2010) compared the applicability of two types of online-recommender systems for mobile in-store usage. They found that users who compared complete products with one another achieved higher decision accuracy than users who compared products along attributes (for example, users they compared products first across their prices and then across their brands). Thus, the authors achieved context-awareness by taking into account the more typical of these two ways of comparing products in-store which is the complete product comparison. Though their approach can be classified as interactive and personalized, the degrees of intuitiveness, attractiveness and social components are limited and their system requires extensive explicit user input.

Other researchers have suggested using location-awareness to build context-aware systems. Kawashima et al. (2006), for example, estimate the user's interest in an object based on the user's physical distance from the objects in the store. Fang et al. (2012) estimate the user's preferences for a brand using the time they spent on a particular brand in a store and how often they look at it. Their system had a higher ease of use, usefulness and satisfaction than a benchmark that required explicit user input. We think that among the systems presented, these context-aware systems that take into account location-awareness come closest to what an UDSS for in-store purchase decisions should look like. However, geospatial location is too coarse to convey relevant information for decisions in a supermarket, where shelves are closely packed with different product types. In the following, we will thus present the AMICA platform that supports more sophisticated and fine-grained localizations, not only of the user, but also of the user's target of visual attention.

11.3 Towards a UDSS-Design

For the design of AMICA, we orient ourselves on the requirements described by Vodanovich et al. (2010) and Junglas and Watson (2006), as they cover most of the requirements found by other research groups (Resatsch et al. 2007, 2008; Tilvawala et al. 2011). Junglas and Watson (2006) identified four fundamental requirements for UDSS in shopping environments: ubiquity, universality, uniqueness, and unison. Ubiquity is defined as reachability, accessibility and portability. Universality refers to universal mobile devices. Uniqueness means that users can be identified and localized. Thus, this concept is similar to personalization. Unison calls for integration of data so that people have a consistent view of information. Vodanovich et al. (2010) suggest a guiding list of requirements particularly for UDSS if they are used by digital natives that includes: (1) personalized, (2) interactive, (3) intuitive (4) attractive and (5) social. Personalization refers to "the ability to provide content and services that are tailored to individuals based on knowledge about their preferences and behaviors" (Adomavicius and Tuzhilin 2005, p. 84). Interactivity is usually defined as the quality of being interactive, i.e. responding to previous actions. Intuitive refers to interfaces that can be navigated without further explanation. The attractive dimension is achieved by including "cool" and up-to-date designs (Vodanovich et al. 2010, p. 719). Social systems allow users to express their own identities or showing users who contributed what.

We try to achieve Ubiquity by putting the UDSS in objects which are commonly used daily: glasses. They are an accepted and often necessary accessory of our culture. Besides their primary function, different manifestations for sports, fashion or safety exist. In contrast to the existing approaches that work with mobile devices, such as smartphones, we believe glasses to be more ubiquitous because they are wearables that users will likely over time integrate into their body schema and as the envisioned system does only require little if any explicit interaction they are much easier to handle. The technical basis of the UDSS, such as the AMICA prototype described in the following, may provide an open platform for many extensions (Apps). Such an UDSS is thus a sophisticated technical device and we expect it to be attractive for users.

Universality is guaranteed by building on top of existing mobile technology which includes means for mobile communication and mobile apps. In addition, we also introduce cognitive apps, applications that are tailored to specific interaction contexts which require little explicit user interaction but are based on cognitive models and are triggered by observing behavioral patterns of the user. In fact, the UDSS described in this paper is only one kind of such cognitive apps. This should also increase attractiveness, as the user is not required to handle additional technologies (cell phone/smartphone).

Uniqueness is given on multiple levels. Glasses are very personal devices with distinct ownership. The system includes different technologies for localization (GPS, WiFi, 3G, NFC, Accelerometer, Gyroscope, and Compass) and thus supports a solid level of context awareness.

The system's cognitive architecture is tailored to support a high-level of personalization, e.g. by adapting to the goals of the user. A conversational interface is at the heart of the system, which adds to personalization and supports a social binding between the user and the system. The possibility to use speech, gaze and gestures to communicate with the system should make it very intuitive.

Unison is supported by a cognitive architecture that supports means-ends reasoning and an elaborated memory model. Interaction with the system is a social activity on its own, but common technologies which make use of social media can easily be integrated as well.

11.4 Attentive Mobile Interactive Cognitive Assistant (AMICA)

AMICA is a platform for personal ubiquitous computing. The underlying architecture is that of an intelligent agent who has the capabilities to perceive its environment, reason about it and act accordingly (Russell and Norvig 1995).

The perception of the system is supported by several sensors. Besides the internal proprioception sensors for localization described above, it supports a microphone, a scene camera and an eye-tracking system (see Figs. 11.1 and 11.2). The use of mobile eye-tracking is a unique feature of the system. It supports a highly localized detection of the visual attention of the wearer and thus enables increases context awareness beyond geospatial localization and basic activity recognition (see Meißner et al. 2013 for a discussion of requirements).

A cognitive architecture based on a belief-desire-intention (BDI) architecture is at the core of the reasoning system (Bratman 1999). It supports modal logic to represent believes about the world and about the intentions and goals of the user. The architecture supports domain-specific extensions, called cognitive apps, for different contexts and daily activities. The idea is that the system dynamically detects the current situation and enables relevant cognitive apps dynamically. For example, in a current prototype we have a cognitive app for chess tutoring which automatically

Fig. 11.1 The AMICA system prototype is based on standard technologies, such as a laptop (*backpack*), microphones, earphones, a scene camera and a binocular eye-tracking system

Fig. 11.2 The AMICA
system in action. The
backpack contains the laptop
hosting all the functionality
of AMICA
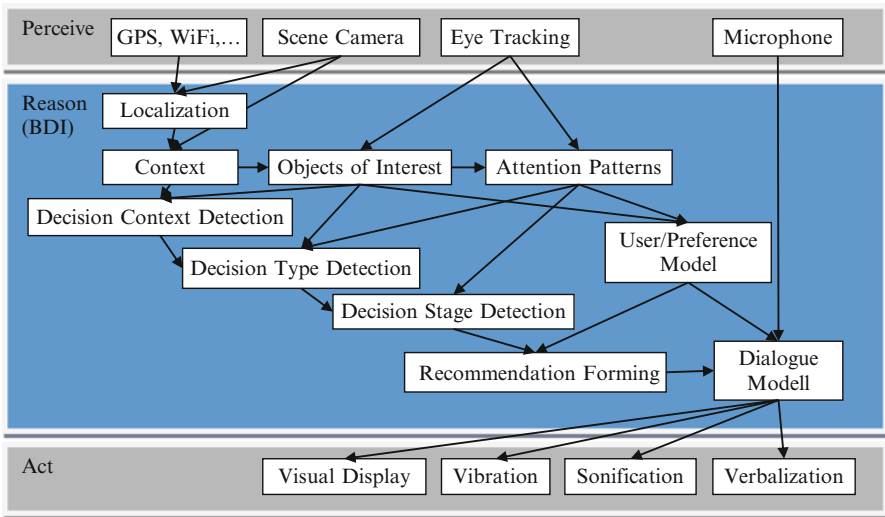
Fig. 11.3 Information architecture for the AMICA system

detects chess boards using computer vision and provides hints to the user to support
them in learning chess. Figure 11.3 displays the information flow of the system's
components. Based on the available sensors, the system localizes itself and detects
the context of the current interaction. If a particular context, e.g. a store, aisle or
shelf, is detected, the decision support system app is activated and a more fine-grained
detection of the current decision context is triggered. The system then continues
with the detection of the decision type that defines the user's intention (e.g., goal-
oriented buying of a particular product versus exploratory browsing, see below) and
then detects the exact stage in the decision process (e.g., orientation, evaluation and
validation; Russo and Leclerc 1994). Based on this detection of the context, recom-
mendations can be communicated in the dialogue with the user taking into account
the model of the user and her preferences.

The current version of AMICA conveys information to the user using audio output (sonification or voice), but extensions to support a near-eye visual display are possible, such as the available EPSON Moverio BT-200 or the upcoming GlassUp or Google Glasses. Using vibration, the user can be subtly made aware of potential decision support to be offered by AMICA.

11.5 Context Detection: Goal-Oriented Versus Exploratory Decisions

One important aspect of AMICA is the unobtrusive detection of the current context and of the tasks the users are occupied with. Geospatial localization can narrow down the set of possible decision contexts. Detecting the current task requires more sophisticated methods. As a unique approach, AMICA tries to infer tasks based on visual information about the current scene, e.g. the location in front of a certain shelf (see Figs.11.4 and 11.5) using computer-vision methods (Harmening and Pfeiffer 2013), combined with the observation of the attentional processes of the user which can be found out through the use of eye-tracking.

In recent years, more sophisticated mobile eye-tracking systems have enabled researchers to investigate attentional processes in natural environments, like supermarkets. It has been shown that attentional processes can differ considerably

Fig. 11.4 Scanpath of participant 4 during the exploratory task
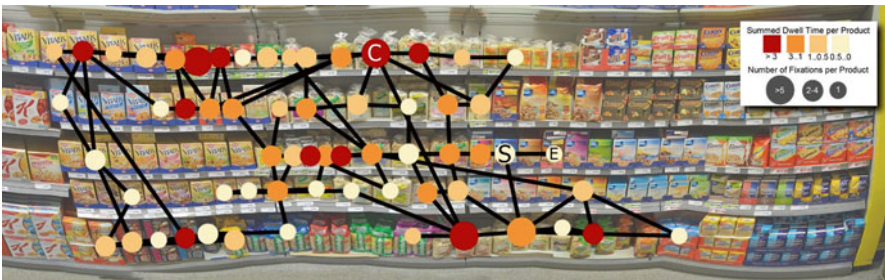
Fig. 11.5 Scanpath of participant 7 during the goal-oriented task

between laboratory settings and more natural environments (see e.g., Hayhoe and Ballard 2005) and that attentional processes are highly task dependent (Gidlöf et al. 2013). Castelhano et al. (2009), for example, found large differences in eye movements when participants processed information under two instruction sets: visual search and memorization.

In line with the above research, we expect similar effects in a supermarket. Imagine, for example, that a consumer has planned to buy muesli. When entering the supermarket, she already knows that she likes mueslis with chocolate and almonds and that the muesli should be low in calories. For this planned purchase, the consumer obviously is goal-oriented. Now imagine another consumer who does not know which characteristics are important for her and therefore is browsing the supermarket aisles. This kind of attentional process can be best described as exploratory. This distinction can be derived from research on search behavior which implies that searching can be dichotomized into goal-directed versus exploratory search (Janiszewski 1998).

We follow the idea by Moe (2003) with the aim to find indicators that allow us to differentiate between goal-oriented versus exploratory tasks in in-store purchase decisions using AMICA. We argue that a UDSS that is able to automatically detect whether a goal or exploratory context is given will help to better adopt the decision support provided to the users' needs. In another study we found that users more strongly prefer to receive ratings and comments and product recommendations in goal-oriented tasks than in exploratory tasks (Pfeiffer et al. 2014). Furthermore, in exploratory task a perfect detection rate of products is more important than in goal-oriented tasks. More work is needed to find out about different requirements for decision-support in different purchase decisions.

In the following, we investigate differences in attentional processes when consumers are manipulated to make goal-oriented and exploratory decisions and observe whether the two cognitive tasks can be identified using attentional processes.

11.6 Evaluation of AMICA

11.6.1 Setup of the experiment

We conducted an experimental study in a medium-sized grocery store. Twenty shoppers were recruited directly after entering the store and they received 10 € as incentive for participation. The mean age was 31.3 (standard deviation (std.) = 13.27, maximum 53 years) and 70 % were female. We chose muesli as product category, because it is information-intensive, offered a sufficient variety of 116 different products, and the packages have a form that can be more easily annotated for the eye-tracking analyses than other products. Furthermore, the packages are more or less of equal size which makes certain measures, such as the distance between products considered, easier to compare. Participants were very different with respect to their interest in muesli. Five reported to never buy muesli, 7 reported to

buy muesli fewer than once a month, 3 once a month and 5 several times per month. Fifteen participants said that they eat muesli up to 3 times per week. The mean of how often muesli is consumed per week is 1.95 (std = 2.26).

The participants were randomly assigned to either a goal-oriented task (GT) or an exploratory task (ET), yielding a group size of ten respondents for each of the two decisions tasks. In each group, they were first read out the task description and then the experimenter ensured that participants had understood the task. In GT they were told to select a muesli for a friend who would come for a visit. In that scenario, the friend likes to have a muesli which fulfills two binary criteria and one continuous criterion, i.e. the muesli should (1) contain chocolate, (2) contain almond, (3) be as low in calories as possible. Eight products fulfilled both binary criteria and there was one optimal product that was the lowest in calories. In ET, participants were asked to gain a fairly good overview about the muesli assortment and to determine criteria which are important for them when buying muesli. Afterwards, they had to choose one product they would potentially buy themselves.

During the task, participants had to wear the AMICA prototype. Running in non-interactive mode, the sensors of AMICA (scene-camera, microphone and eye-tracking system) recorded the behavior of the participants, but the system provided no feedback. This way, we collected first data on the acceptance of wearing such a system in public areas and on the benefit of envisaged functionalities with a post-hoc interview, while at the same time, we were able to record a corpus of real-life data on customer behavior in specified contexts with identified intentions.

11.6.2 Results: Differentiating Between Decision Situations Based on Eye Movements

Table 11.1 summarizes the values for describing the decision processes of respondents in the GT versus ET: Four observations are missing because of technical problems with the USB-port during recording of the eye movements. The results show that

Table 11.1 Key differences between the performances in GT and ET

	GT (n=9)	ET (n=7)
a) Length of task	192.22 s	65.52 s
b) Average distance between two fixations on different products	33.67 cm	45.78 cm
c) Number of different products looked at	58.67	45.38
d) Fixations on products including re-fixations on same products	131.89	76.25
e) Time looked at one product (summed up over re-fixations)	2.44 s	0.84 s
f) Length of continuous product fixations	1.28 s	0.54 s
g) Time spent on the last three products considered	18.56 s	7.81 s
h) Time looked at brand and logo	84.88 s (59.23 %)	29.05 s (80.36 %)
i) Time looked at price	2.31 s (1.61 %)	3.24 s (8.96 %)
j) Time looked at detailed information	56.11 s (39.16 %)	3.86 s (10.68 %)

several measurements are potentially useful for detecting whether a user is in a goal-oriented or exploratory decision situation. In the GT, respondents put much more effort in the purchase decision (a,c,d,g). They made more fixations (d) on more products (c) and subsequent fixations were closer together (b). They also spent more time (e) on individual products, especially before making their decision (g). This last result is in line with Gidlöf et al. (2013) who found an attentional focus to the finally chosen product when participants selected products from a supermarket shelf. We categorized information available on muesli packages in the following three categories: h) Brand and logo, i) price and j) detailed information, such as ingredients. In the GT, information about brand and details were important while it was primarily brand for the ET.

These behavioral differences also show up in typical scanpaths of participants for ET (Fig. 11.4) and GT (Fig. 11.5). Each fixation is represented by a circle and eye-movements between fixations are drawn as paths between these circles. Larger circles represent more fixations on the same object, more intense colors represent a longer summed dwell time. S stands for start (the first fixation) and E for end (the last fixation).

These first results show that when exploring information in a product category, participants focus on brand-related and price information and neglect detailed information. However, when pursuing the goal to select a product based on predefined preferences, participants acquire more detailed information. In sum, the results show that the level of processing of detailed information can provide information about the goal orientation of the participant. The aim of the UDSS therefore should be to identify the degree of goal-orientation and, based on that, to give detailed information which is of help in the purchase situation at hand. The next aim of our project therefore is to investigate how a recommender system can be trained to learn which specific information is relevant in a certain decision situation.

11.7   Conclusion

The design of mobile recommender systems for in-store usage gives rise to several research questions. We are developing such a system based on the presented AMICA platform. In this paper, we required ground-truth data on real-life attentional processes during decision making in front of a supermarket shelf. The AMICA system provides advanced technologies for context-awareness based on computer vision using a scene camera and eye-tracking. This promises unmatched possibilities to monitor the attentional processes of the user in real-time and enables us to differentiate between decision tasks with distinct needs of support from a recommender system. The results of our study show that the attentional processes during the goal-oriented and the exploratory tasks bear enough differences in basic eye-movement features, such as duration and number of fixations, or the ratio of fixations on detailed information versus brand and logo, in order to allow us to differentiate between the two tasks.

These first results suggest that using the AMICA framework for mobile recommender systems is promising. In the future, we would like to build a classifier using an appropriate subset of indicators that describes the difference in decision processes between goal-oriented versus exploratory tasks. Next, we would like to investigate the differences in the users' needs in the two different decision contexts and define the decision model. Putting both results together, we should be able to implement a first prototype of AMICA.

Biography

Jella Pfeiffer is Akademische Rätin (assistant professor) in Information Systems and Business Administration at the Karlsruhe Insititue of Technology (KIT), Germany. Her main research interest is in recommender systems. Further research interests are in decision aids in e-commerce, consumer decision behavior and social networks. Jella holds a Doctoral degree in Information Systems from the Johannes Gutenberg-University Mainz and has been a research fellow at Harvard, University of British Columbia and University of Lausanne.

Thies Pfeiffer is Akademischer Rat (assistant professor) at the Center of Excellence "Cognitive Interaction Technology" (CITEC) of Bielefeld University, Germany. His main research interest is in multimodal interaction in 3D worlds, with a special focus on gaze and gestures. Further research interests are interaction techniques for virtual reality, virtual agents, scientific visualization and social networks. Thies is one of the organizers of the workshop on Solutions for Automatic Gaze-Data Analysis held in 2013 at CITEC (http://saga.eyemovementresearch.com). He received his Doctor rer. nat. from Bielefeld University in 2010 on the interaction of gaze and gestures in deictic reference.

Martin Meißner is an associate professor of marketing at the University of Southern Denmark. His main research interest is in understanding the relationship between preference and attention by recording eye movements and applying other process tracing techniques. Martin has published one of the first studies on eyetracking in discrete choice experiments and holds a grant from the German Research Foundation (DFG) to study decision sequences using eyetracking. Martin holds a Doctoral degree from Bielefeld University and has been a visiting fellow at Duke University and at Monash University

References

Adomavicius, G., & Tuzhilin, A. (2005). Personalization technologies: A process-oriented perspective. *Communications of the ACM: The digital society, 48*(10), 83–90.

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 217–253). Berlin: Springer.

Adomavicius, G., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems, 23*(1), 103–145. New York, USA.

Ansari, A., Essegaier, S., & Kohl, R. (2000). Internet recommendation systems. *Journal of Marketing Research, 37*(3), 363–375.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research, 25*(3), 187–217. http://www.jstor.org/stable/10.1086/209535.

Bratman, M. E. (1999). *Intentions, plans, and practical reasoning*. Stanford: CSLI.

Castelhano, M. S., Mack, M. L., & Hendersson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*, 1–15.

Fang, B., Liao, S., Xu, K., Cheng, H., Zhu, C., & Chen, H. (2012). A novel mobile recommender system for indoor shopping. *Expert Systems with Applications, 39*(15), 11992–12000.

Gidlöf, K., Wallin, A., Dewhurst, R., & Holmqvist, K. (2013). Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment. *Journal of Eye Movement Research, 6*(3), 1–14.

Harmening, K., & Pfeiffer, T. (2013). Location-based online identification of objects in the centre of visual attention using eye tracking. In *Proceedings of the first international workshop on solutions for automatic Gaze-data analysis 2013 (SAGA 2013)* (pp. 38–40). Bielefeld, Germany.

Häubl, G., & Murray, K. B. (2003). Preference construction and persistence in digital marketplaces: The role of electronic recommendation agents. *Journal of Consumer Psychology, 13*(1–2), 75–91.

Hayhoe, M. M., & Ballard, D. H. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*, 188–194.

Janiszewski, C. (1998). The influence of display characteristics on visual exploratory search behavior. *Journal of Consumer Research, 25*, 290–301.

Junglas, I. A., & Watson, R. T. (2006). The U-constructs: Four information drives. *Communications of the Association for Information Systems, 17*(1), 2–43.

Kawashima, H., Matsushita, T., Satake, S., Imai, M., Shinagawa, Y., & Anzai, Y. (2006). PORSCHE: A physical objects recommender system for cell phone users. In *Proceedings of 2nd international workshop on personalized context modeling and management for UbiComp applications*. California, USA.

Lee, Y. E., & Benbasat, I. (2010). Interaction design for mobile product recommendation agents: Supporting users' decisions in retail stores. *ACM Transactions on Computer-Human Interaction, 17*, 4.

Meißner, M., Pfeiffer, J., & Pfeiffer, T. (2013). Mobile eyetracking for decision analysis at the point-of-sale: Requirements from the marketing research and human-computer interaction-perspective. In *Proceedings of the 1st international workshop on solutions for automatic Gaze Data analysis (SAGA 2013)* (pp. 24–25). Bielefeld, Germany.

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology, 13*(1&2), 29–39.

Murray, K. B., Liang, J., & Häubl, G. (2010). ACT 2.0: The next generation of assistive consumer technology research. *Internet Research, 20*(3), 232–254.

Palmisano, C., Tuzhilin, A., & Gorgoglione, M. (2008). Using context to improve predictive modeling of customers in personalization applications. *IEEE Transactions on Knowledge and Data Engineering, 20*(11), 1535–1549.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press, Cambridge, UK.

Pfeiffer, J., Huschens, M., & Pfeiffer, T. (2014). Important product features of mobile decision support systems for in-store purchase decisions: A user-perspective taking into account different purchase situations. In *Proceedings of the Multikonferenz Wirtschaftsinformatik (MKWI)*. Paderborn, Germany.

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the Horizon, 9*(5), 1–6.

Resatsch, F., Sandner, U., Michelis, D., Hoechst, C., & Schildhauer, T. (2007). Everyday simplicity: The implications of everyday tasks for ubiquitous computing applications. *Americas Conference on Information Systems* (p. 105). Keystone, Colorado, USA.

Resatsch, F., Sandner, U., Leimeister, J. M., & Krcmar, H. (2008). Do point of sale RFID-based information services make a difference? Analyzing consumer perceptions for designing smart product information services in retail business. *Electronic Markets, 18*(3), 216–231.

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs: Prentice-Hall.

Russo, J. E., & Leclerc, F. (1994). An eye-fixation analysis of choice processes for consumer nondurables. *The Journal of Consumer Research, 21*(2), 274–290.

Slovic, P. (1995). The construction of preference. *American Psychologist, 50*(5), 364–371.

Tapscott, D. (2008). *Grown up digital: How the net generation is changing your world*. New York: McGraw-Hill.

Tilvawala, K., Myers, M., & Sundaram, D. (2011). Design of ubiquitous information systems for digital natives. In *Proceedings of the Pacific Asia conference on information systems*. Brisbane, Australia

van der Heijden, H. (2006). Mobile decision support for in-store purchase decisions. *Decision Support Systems, 42*(2), 656–663.

Vodanovich, S., Sundaram, D., & Myers, M. (2010). Digital natives and ubiquitous information systems. *Information Systems Research, 21*(4), 711–723.

von Reischach, F., Michahelles, F., Guinard, D., Adelmann, R., Fleisch, E., & Schmidt, A. (2009). An evaluation of product identification techniques for mobile phones. In *Proceedings of the 12th IFIP TC 13 international conference on human-computer interaction: Part I* (pp. 804–816). Uppsala, Sweden.

Chapter 12
Engaging with Online Crowd:
A Flow Theory Approach

Cuong Nguyen, Onook Oh, Abdulrahman Alothaim,
Triparna de Vreede, and Gert Jan de Vreede

Abstract Online collaborative problem solving (OCPS) refers to the use of social web technologies to garner netizens' collective effort for problem solving and innovation tasks. The model has enabled organizations to involve online users in organizational works at large scale. However, success of this kind of initiatives depends much on, among other things, user engagement, or the amount of effort online users voluntarily devote to what are requested in an OCPS initiative. We argue that an important influence on user engagement in OCPS events is their experience when participating in the events. We further argue that Flow Theory by Csikszentmihalyi and Csikszentmihalyi (1988) provides much insights on how to improve this experience. In addition, we propose to measure the psychological construct "flow" through a novel physiological-psychometric approach. In this paper, detailed discussion of our theoretical standpoint and the design of a lab experiment to validate our hypotheses are provided.

Keywords Crowdsourcing • Online collaborative problem solving • Flow

12.1 Introduction

Social web technologies offer unprecedented opportunities for organizations to 'crowdsource' tasks to a large number of online users (Howe 2006). One of the strategies that organizations often employ in their crowdsourcing projects is online

C. Nguyen (✉) • A. Alothaim • T. de Vreede • G.J. de Vreede
The Center for Collaboration Science, University of Nebraska at Omaha,
6708 Pine Street, MH 399D, Omaha, NE 68182, USA
e-mail: cdnguyen@unomaha.edu; aalothaim@unomaha.edu;
tdevreede@unomaha.edu; gdevreede@unomaha.edu

O. Oh
Information Systems Group, University of Colorado Denver Business School,
1475 Lawrence St. Denver, Co 80208, USA
e-mail: onookoh@gmail.com

collaborative problem solving (OCPS). In an OCPS project, organizations invite online users to discuss a problem or issue through a social web platform. The model is similar to that of traditional offline meetings with the exception that OCPS meetings are held asynchronously among an undefined number of people with a frequent change of membership or attendance. A typical example of an OCPS platform is MindMixer (mindmixer.com). In MindMixer, a civic problem is posted by a government agency (e.g. "How to make Omaha a better city to live") so that online users can (1) suggest as many ideas as possible through online posts, and (2) evaluate each suggested idea by commenting, voting, or rating to consolidate the large number of suggested ideas into a best few ideas worthy of more attention by the government agency.

The OCPS model is often praised for having advantages in: (1) accessing diverse ideas for innovative problem solving, (2) detecting customer needs, or (3) in the public domain, increasing citizens' commitment to policy changes (Aitamurto et al. 2011; Bommert 2010). These advantages are based on the assumption that OCPS platforms can attract large numbers of online users to solve organizations' problems. Despite the desired advantages, however, convincing people to work on somebody else's problems through the Internet is a challenging issue (Brabham 2009; Doan et al. 2011). Therefore, our research question is: "What are the factors that enhance the level of online user engagement to solve organizations' problems?"

In this study, online user engagement is defined as the amount of effort online users voluntarily devote to what are requested in an OCPS initiative. This definition emphasizes the behavioral manifestation of engagement rather than its cognitive or emotional aspects (Appleton et al. 2008; Schaufeli et al. 2002). Our preference for behavioral engagement stems from the fact that to organizers of crowdsourcing events, online user engagement is significant only when they actually contribute meaningfully to an event.

For this study, we adopt Flow Theory (Csikszentmihalyi and Csikszentmihalyi 1988), a theory on human optimal experience. We argue that compared to other research on user motivation in crowdsourcing, Flow Theory provides more practical insights to practitioners. We will conduct a lab experiment to test our hypotheses in a scenario chosen from the actual OCPS platform Mindmixer.com. Moreover, in response to Dimoka et al. (2012) call for applying physiological measurements of psychological constructs, we will measure the psychological construct "flow" through a combination of psychometric measurement scales and eye tracking.

The remainder of this paper is organized as follows. In the "Theoretical Background" section, we introduce the Flow Theory and highlight its utilities in comparison to previous literature on the crowdsourcing user motivations. We will also present the rationale for combining physiological with survey measurements of the "flow" construct. Next, we describe our lab experiment plan in the "Research Method" section. We conclude the paper with a discussion of expected contributions of this work.

12.2 Theoretical Background

12.2.1 Previous Research on User Motivation in Crowdsourcing

Reflecting the importance of user engagement in crowdsourcing, many research attempts have been made to understand why online users participate and contribute in crowdsourcing events (e.g. Boudreau and Lakhani 2009; Borst 2010; Brabham 2012; Kaufmann et al. 2011). These studies, in general, classify the drivers of online users' engagement into two types: (1) the activity by itself is interesting to the online users (intrinsic motivation), or (2) they perform the activity because of a specific reward (extrinsic motivation) (Ryan and Deci 2000). A number of intrinsic motivators have been identified, including fun, fame, professional and personal identity or altruism (Boudreau and Lakhani 2009; Brabham 2012; Wagner and Prasarnphanich 2007). Moreover, online users' exertion of time and effort for crowdsourcing projects can be driven by extrinsic motivators such as money, social pressure, peer support, career advancement or skill development (Boudreau and Lakhani 2009; Brabham 2012; Malone et al. 2009).

For OCPS initiatives, in particular, previous studies indicate that user participation is more driven by intrinsic rather than extrinsic motivators (Boudreau and Lakhani 2009; Lakhani and Wolf 2005; Wagner and Prasarnphanich 2007; Zheng et al. 2011). That is, whether online users are willing to work in an OCPS initiative depends on how interesting the crowdsourcing task is to them, not on what they can get in exchange for their contribution of time and effort. Thus, to achieve high user engagement in OCPS initiatives, the crowdsourcing task itself should match the users' individual preference and interests.

The fact that online users may be more intrinsically motivated to participate in collaborative problem solving model is valuable to practitioners, but does not provide sufficient guidance to actually create interventions to increase user engagement in crowdsourcing events. Interestingness seems to depend much on the nature of the task and individual preferences. For example, an online user participates in an open source software project because he/she likes programming. Therefore, it is less than ideal to apply this insight to other context, such as contributing ideas to improve municipal administrative works, because the tasks stem from different contexts. Therefore, further insight should be provided from two questions: (1) *what factors makes an activity interesting?* and (2) *what factors can be created or manipulated to make activities more interesting and engaging?*

12.2.2 Flow Theory

A key to the two issues raised above can be found in Flow Theory by Csikszentmihalyi and Csikszentmihalyi (1988). Csikszentmihalyi and colleagues found that individuals, when working or playing, can reach an optimal experience where they find the

activities they are doing were rewarding in and of itself regardless of the activity's outcomes (Csikszentmihalyi 1975). Csikszentmihalyi called this experience, *flow*, a word often mentioned by his subjects when they were asked to reflect on the experience. Csikszentmihalyi (1990) defines flow as "the state in which people are so involved in an activity that nothing else seems to matter; the experience itself is so enjoyable that people will do it even at great cost, for the sheer sake of doing it."

The existence of flow have been confirmed later on by empirical studies in various domains such as music and arts (Bakker 2005), sports (Jackson et al. 2001), media (Sherry 2004), work and life quality (Eisenberger et al. 2007), and online activities (Hoffman and Novak 2009). Flow was also reported to lead to positive consequences such as increased performance, increased learning and exploratory behaviors, or intention to return (Eisenberger et al. 2007; Engeser and Rheinberg 2008; Hoffman and Novak 2009).

Beyond discovering the flow experience, Csikszentmihalyi also identified the three external conditions that usually relate to the occurrence of the flow experience (Csikszentmihalyi and Csikszentmihalyi 1988). First, the person's skill needs to match the challenge the activity posed (i.e. *challenge-skill balance*). If his/her skill exceeds the challenge, (s)he would get bored. If his/her skills are below the challenge, (s)he would feel anxiety. Second, the activity needs to have a *clear goal*. A person needs to know what (s)he is aiming for. Finally, the activity should provide *immediate and unambiguous feedback* so that a person had a clear sense of progress and performance. The role of the three conditions as precursors of flow was later supported by a number of lab experiment and field studies (e.g. Engeser and Rheinberg 2008; Guo and Poole 2009; Keller and Bless 2008; Mannell and Bradley 1986).

Within our study context, Flow Theory addresses both aspects of an activity's interestingness mentioned above. An interesting activity is one for which individuals can experience a flow state. However, while other intrinsic motivators (e.g. fun, fame, personal identity) depend much on a person's internal traits and preferences, a flow state emerges from the interaction between the person and his/her surrounding environment (Nakamura and Csikszentmihalyi 2002). To illustrate, we can look at motivation to play badminton. A person's prior perception that badminton is a fun and wholesome sport might be a driver to play it. However, once (s)he starts playing it, the flow experience of playing badminton itself may drive the person to keep playing it for hours.

The fact that a flow state emerges through interaction between a person and his/her surrounding environment is significant to practitioners. While intrinsic motivators typically depend on a person's internal traits and preferences, the flow state depends on external factors. These external factors, in turn, can be manipulated. That is, practitioners can develop interventions to stimulate the flow state. Moreover, the fact that flow exists in various activities including both work and leisure (Csikszentmihalyi and Csikszentmihalyi 1988) indicates that flow does not depend on *what the task is, but how the task is structured*.

Applying Flow Theory to the OCPS context, we expect that by structuring the OCPS tasks to meet the three flow conditions (a clear goal, challenge-skill balance, and immediate and unambiguous feedback), online users can experience a flow
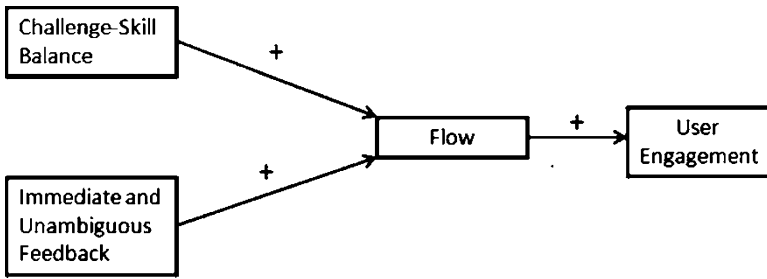
Fig. 12.1 Theoretical model

state when they engage in the event and subsequently contribute more to the event. Furthermore, we argue that while clear goal is naturally a must for guiding people working on OCPS tasks, the effects of challenge-skill balance and feedback on user engagement are not so obvious. Therefore, within this specific study, we treat "clear goal" as a control variable and propose to examine the following propositions through a lab experiment:

In the online collaborative problem solving context:

P1: A challenge-skill balance positively relates to the invocation of the flow state.
P2: Immediate and unambiguous feedback positively relates to the invocation of the flow state.
P3: A flow state positively relates to user engagement.

The theoretical model is illustrated in the Fig. 12.1.

12.2.3 Measuring the Flow State: A Physiological-Psychometric Approach

To measure the flow experience, we propose to use a novel approach that combines the use of a survey and eye tracking tool. In fact, how to measure the flow state has been a thorny problem (Finneran and Zhang 2005; Hoffman and Novak 2009; Moneta 2012). To date, dominant measurement methods include a survey and the experience sampling method (ESM) (Moneta 2012). The survey method typically asks subjects to reflect on their flow experience in an activity by rating their experience on a Likert scale or by answering some open-ended questions (Chen et al. 1999; Hoffman and Novak 2009). While those methods are easy to conduct (Hoffman and Novak 2009), they have limitations in that they can only capture subjects' conscious impressions on their flow experiences, which might be distorted over time and inaccurate (Finneran and Zhang 2005). To complement this limitation, ESM has been used where the subjects are asked to fill out short questionnaires at specific intervals while they are performing the activity under study (Csikszentmihalyi 1990). However, this method runs the risk of interrupting the flow state of the subjects (Finneran and Zhang 2005).

In this study, we propose combining a survey with physiological approach to measure the flow state. Specifically, we will use an eye-tracking tool, a device that keeps track of eye movements, to capture the flow experience. Physiological methods to measure psychological constructs has recently gained popularity and has been recommended for IS research (Dimoka et al. 2012). It has a number of advantages over surveys and ESM. Rather than relying on *subjective* reflection of the subjects on their flow experiences, the physiological approach can capture *objective* indicators in *real time* by recording eye gaze and movement. The real time measuring of eye movements is important to a construct like flow that can vary over time. In addition, the fact that eye gaze and movement are automatically tracked through sensors provides an unobtrusive approach to collect measurement data.

However, a key issue with physiological approach is to explicate how the observed physiological indicators are associated with the psychological and latent construct. Specifically, a key question is to clarify how eye movements are related with the flow state of research subjects. Csikszentmihalyi and Csikszentmihalyi's (1988) assumptions regarding human consciousness can answer this question. To explain Flow Theory, Csikszentmihalyi and Csikszentmihalyi (1988) assumed that humans have a cognitive mechanism called *consciousness* to process external stimuli. The consciousness is a system composed of three components: *attention, awareness,* and *memory.* Attention receives external stimuli and transfers them to the awareness. The awareness interprets the stimuli and accordingly triggers cognitive and affective responses. The awareness also decides whether to store them in the long-term memory.

The interesting part of this mechanism is that the consciousness is considered to be a limited resource to process attention. That is, the attention can process only a limited number of stimuli at a time. When a task challenge matches a person's skills, the stimuli created by the challenge exhaust this limited attention resource and the person can only focus on the challenge and on nothing else. If the task challenge is below a person's skill, the person can have spare attention resources to attend to other things and may consequently lose concentration. If the challenge is above a person's skill, (s)he will feel stressed.

This mechanism implies that when a person is in a flow state, his senses (body components that are in charge of receiving external stimuli) should all be occupied by stimuli relevant to the task at hand. In the OCPS context, the online users perform activities (e.g. provide contributions) through their interaction with the user interface (UI) of the crowdsourcing platform. Therefore, it can be assumed that the relevant stimuli to the contribution acts should come from the system's UI. Consequently, we assume that an indicator of a person's flow state is the extent to which *his/her eye gaze is focused only on the UI features created to support the activity* under study.

While Flow Theory theoretically supports an association between eye movement and flow state, an apparent problem with using eye-tracking data is that they do not provide an access to participants' thoughts (Eger et al. 2007). For example, eye-tracking data alone is insufficient to determine if a participant's fixated gaze is due

to the subject's high cognitive load or difficulty in processing information. Therefore, in addition to eye-tracking, a survey with traditional psychometric measures of flow experience will also be used to capture the explicit reflection of the participants on their flow experience.

12.3   Research Method

In this study we will test our theoretical model through a lab experiment. The experiment aims at determining the effect of two independent variables (IVs) on the flow state and user engagement. These IVs are challenge-skill balance and immediate and unambiguous feedback. The IV challenge-skill balance has three values "challenge=skill", "challenge>skill", and "challenge<skill". The IV immediate and unambiguous feedback has two values "enabled" or "not enabled". Therefore, structurally, this is a 3×2 factorial design with six treatments in total. To ensure the statistical power of 0.8 as recommended in (Baroudi and Orlikowski 1989) and relevance of the subjects to the experiment task, a minimum of 90 students from the University of Nebraska at Omaha (UNO) will be recruited. This sample size is based on the output of GPower, an a priori power analysis software tool (Faul et al. 2007).

Experiment subjects will be asked to brainstorm ideas on the topic "How to make UNO a better university" through a web page specifically designed for this experiment. The task is chosen such that it is both similar to real tasks in Mindmixer (www.mindmixer.com) OCPS projects and suitable for the experiment subjects.

12.3.1   Operationalization of the Independent Variables (IVs)

12.3.1.1   Challenge-Skill Balance

To enable different values of this IV, we will use a set of cognitive challenges related to the topic "How to make UNO a better university" with three difficulty levels. The difficulty levels of the challenges are created based on the revised version of Bloom's taxonomy of learning objectives (Krathwohl 2002). According to this taxonomy, there are five types of cognitive process (listed in the increasing level of sophistication) as *remember, understand, apply, analyze, evaluate,* and *create*. Following this classification, we created the challenges such that higher level of challenge difficulty invokes more sophisticated cognitive process. In addition, to ensure that the level-3 challenges are indeed very challenging, we request the respondents to address specific and hard problems, as opposed to letting them choose any issues they like to talk about in level-2 challenges. Description of the cognitive process invoked at different levels is provided in the Table 12.1 below, along with the example challenges:

Table 12.1 Three difficulty levels of challenges

Level	Required cognitive process	Example challenges
1	Remember	List things you do on a typical day on the UNO campuses
		List things your friends do on a typical day on the UNO campuses
2	Analyze, apply, evaluate, create	Think of something other universities have that UNO should apply. Justify why it is suitable for UNO
		Think of something that UNO should stop continue doing. Justify why it is so
3	Analyze, apply, evaluate, create (specific and hard problems)	In the effort to build its brand, UNO wants to apply "word of mouth" strategy on social media platforms. For example, UNO would pay people who spread good words about the school on Facebook. However, online users refuse to cooperate because their friends will think that they use them to make money. Propose your ideas on how to overcome this problem

The values of the IV "challenge-skill balance" are manipulated as follows:

Challenge = skill

In this condition, the subjects will be asked to respond to challenges in increasing level of difficulty. Each challenge will last for 5 min. After every 5 min, the subjects will be prompted to work on a new challenge. After every two challenges, the challenge difficulty increases one level. In total, a subject will work on six challenges.

At the end of the session, the subject will be shown the six challenges he/she have worked on and be asked their perceived challenge-skill balance. He/she will also be asked to rate his/her flow state on the time he/she works on those challenges.

Challenge > skill

This condition is enabled by letting the subjects work with the level-3 challenges only. Other manipulations are the same as the "challenge=skill" condition.

Challenge < skill

This condition is enabled by letting the subjects work with the level-1 challenges only. Other manipulations are the same as the "challenge=skill" condition.

12.3.1.2   Immediate and Unambiguous Feedback

In the "feedback is enabled" condition, the participants will receive both quantitative and qualitative feedback on their performance as follows:

- The number of ideas the subjects have submitted will be shown during the experiment session alongside a fake average number of ideas of "all participants". This allows the subject to have relative comparison between his/her performance and other participants' performance.

- Qualitative feedback on the subject's contributions is provided through a notification box where:
  - After the subject submits five ideas, or at the minute 10th, it shows "The review panel is looking at your contributions and says your contributions are relevant. Good job!"
  - At the minute 15th, it shows "The review panel says they are intrigued by some of your contributions".
  - At the minute 22nd, it shows "The review panel says they have identified 3 of your ideas to be recommended to the Chancellor".

  In the "feedback is not enabled" condition, this information is not shown.

12.3.2 Experiment Procedure

Subjects will be randomly assigned to one of the six treatment groups. The six treatments differ in whether the computer user interface enables/constrains the flow conditions (i.e. the IVs' values). Each participant uses a computer to perform the experiment task for 30 min. At the end of the experiment session, the participant is requested to fill out the survey for their flow experiences and other control variables. During the experiment, the participants' eye movements are tracked and recorded. The eye-tracking tool to be used is from Eyetech Digital Systems (http://www.eyetechds.com/usage-research.shtml).

12.3.3 Control Variables

To enhance the internal validity of the experiment design, the variables to be statistically controlled for are task interest, Murphy et al. (1989), education level (e.g., freshman, senior etc.), study major, age and gender.

12.3.4 Measurement of the Observed Variables

12.3.4.1 User Engagement Measurement

By definition, user engagement refers to the amount of effort online users voluntarily devote to what are requested in an OCPS initiative. In the context of this experiment, it is operationalized by the *number of words* of all the relevant ideas a subject generates within the experiment session. The number of words of the ideas, instead of the number of ideas, is used because the higher level challenges in the experiment might require more text to answer properly. The relevance of the ideas to the

experiment task will be evaluated by a group of trained raters on two criteria:
(1) whether the ideas directly addressed the challenges and (2) the ideas are not
verbatim copies of one another.

12.3.4.2 Flow State Measurement

The flow state is evaluated through self-report measurement and eye movements. For
self-report measurement, the participants will fill in an adapted version of the Flow
Short Scale (Martin and Jackson 2008) at the end of the experiment session. For the
eye-tracking, as implied by Flow Theory, the indicator of the flow state is that a per-
son's eye gaze mainly focuses on the external stimuli relevant to the task at hand. In
the experiment, the stimuli are the UI features that are intentionally created for the
brainstorming task, including the instruction statement, the text box to enter ideas,
the idea list, the challenge box, and the status box (see Fig. 12.2). In eye-tracking
terms, these regions are called Areas of Interest (AOI) (Holmqvist et al. 2011).
Data on the following eye-tracking measures will be recorded: dwells, transition
matrices and AOI hits. A dwell records the amount of time the subject's eyes are
within an AOI (Holmqvist et al. 2011). Transition matrices can keep track of eye
transitions between AOIs and areas which are not of interest (Holmqvist et al. 2011).
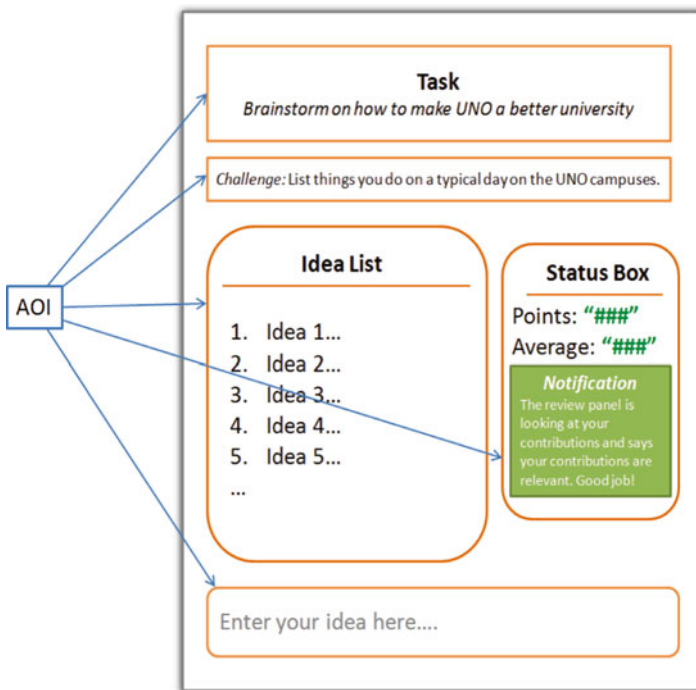
Fig. 12.2 Experiment mock-up screen

For example, eye movements from the idea list to a white space on the screen. Finally, AOI hits count the number of eye fixations (e.g. eyes staring at a place for a relatively long time period) on an AOIs (Holmqvist et al. 2011). It is expected that the flow state is in proportion to the ratio of the total time of all dwells over the experiment duration; in reverse proportion to the eye transitions between AOIs and areas which are not of interest; and in proportion to the ratio of the AOI hits over the total number of eye fixations within the experiment period.

12.4 Data Analysis

Analysis of variance (ANOVA) will be used to examine the main effects and interaction effects of the two IVs on the flow state and user engagement. Moreover, PLS analysis will be applied to measure the proposed theoretical model, which includes measurement validation (reliability testing, convergent and discriminant validity testing), calculation of path coefficients and R-square among the latent constructs, and mediation effect test. Control variables will also be included in PLS analysis to check for confounding effects on the observed causal relationships.

12.5 Conclusions

Crowdsourcing in general and OCPS in particular is a promising model for organizations to tap into the wisdom of online crowds to solve problems. However, to make the model fruitful in practice, the key condition is to attract and retain a large number of contributors. Towards that end, we suggest Flow Theory may provide useful guidance to design an engaging crowdsourcing experience. We plan to conduct a laboratory experiment to test whether the presence of challenge-skill balance, and immediate and unambiguous feedback lead to the flow state and finally to higher user engagement while interacting with an OCPS application.

Our work calls for more attention to the influence of online user experience on their engagement. That is, we suggest that online users will contribute more to the OCPS events if they enjoy their experience during these events. This is consistent with previous research that found that fun is an important motivator (e.g. Boudreau and Lakhani 2009; Brabham 2009; Brabham 2012). However, we argue that it is important to determine the source of this fun. For example, a person might participate in an open source software project because (s)he enjoys programming in general, but also because of their experience when participating in that specific project. In the context of a crowdsourcing event, the former would relate to the users' personal traits, while the later depends on the design and presentation of the crowdsourcing event on the platform. In this study, we recognize this difference and emphasize the later view, with the belief that it will bring significant value to the IS domain.

Furthermore, we expect our work to also make a contribution to IS research methods. Dimoka et al. (2012) calls for the use of physiological tools besides traditional psychometric measurements to capture more well-rounded information on latent constructs. In this study, we argue that this approach is appropriate to measure the flow state and will combine eye-tracking with survey measurement of this construct. We expect that our study will provide a useful example for this emerging research method.

Finally, we expected our study to offer useful insights for practitioners. The three conditions for users to get into a state of flow provide specific guidelines for practitioners to create enjoyable user interface designs for online users to engage in OCPS activities.

Biography

Cuong Nguyen   is currently a Ph.D. candidate in the Ph.D. in IT program of the University of Nebraska at Omaha (UNO). He has also been a research associate of the UNO Center for Collaboration Science since 2008. His research interests include crowdsourcing, collaboration processes, and online engagement.

Onook Oh   is an Assistant Professor in the University of Colorado Denver Business School. He was a research associate at the Center for Collaboration Science in the University of Nebraska Omaha, and an Assistant Professor of the Warwick Business School. He is interested in theorizing the information systems infrastructure, its techno-social impacts on our everyday lives, and technology mediated collective sense-making under crisis situations. He published his researches in MIS Quarterly, Communications of the AIS, AIS Transactions on Human-Computer Interaction, and Information Systems Management, among others.

Abdulrahman Alothaim   is currently a Ph.D. student in the Information Technology program of the University of Nebraska at Omaha (UNO). He has been a research associate at the UNO Center for Collaboration Science since 2012. He is also a teaching assistant in the college of computer and information sciences at King Saud University. His current research interests include social media, crowdsourcing, and collaboration processes.

Triparna de Vreede   is currently completing her PhD in Industrial/Organizational Psychology and pursuing her Master's in Counseling at the University of Nebraska at Omaha. She also holds Master's degree in Management Information Systems and in Business Administration. She is a Research Associate at the university's Center for Collaboration Science. She is a trained facilitator of thinkLets-based Group Support Systems workshops. Her research focuses on the psychology of crowdsourcing, psychological foundations of thinkLet-based collaboration processes, cognitive processes of creativity, technology and work practice transition in groups, and creativity in groups.

Gert Jan de Vreede   is a professor at the University of Nebraska at Omaha and the Managing Director of the university's Center for Collaboration Science. His research focuses on social and organizational applications of collaboration technologies, the theoretical foundations of collaboration, Collaboration Engineering, and the facilitation of group work. He is co-founder of the Collaboration Engineering field and co-inventor of the thinkLets concept. He has published over 250 refereed journal articles, conference papers, and book chapters and was named the most productive GSS researcher world-wide from 2000-2005 in a comprehensive research profiling study.

References

Aitamurto, T., Leiponen, A., & Tee, R. (2011). The promise of idea crowdsourcing–benefits, contexts, limitations. *Nokia Ideasproject White Paper*.

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools, 45*(5), 369–386.

Bakker, A. B. (2005). Flow among music teachers and their students: The crossover of peak experiences. *Journal of Vocational Behavior, 66*(1), 26–44.

Baroudi, J. J., & Orlikowski, W. J. (1989). The problem of statistical power in MIS research. *MIS Quarterly, 13*(1), 87–106.

Bommert, B. (2010). Collaborative innovation in the public sector. *International Public Management Review, 11*(1), 15–33.

Borst. (2010). *Understanding crowdsourcing: Effects of motivation and rewards on participation and performance in voluntary online activities.* PhD thesis, Erasmus University Rotterdam.

Boudreau, K. J., & Lakhani, K. R. (2009). How to manage outside innovation. *MIT Sloan Management Review, 50*(4), 69–76.

Brabham, D. C. (2008). Moving the crowd at iStockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application. *First Monday, 13*, 6.

Brabham, D. C. (2009). Crowdsourcing the public participation process for planning projects. *Planning Theory, 8*(3), 242–262.

Brabham, D. C. (2012). Motivations for participation in a crowdsourcing application to improve public engagement in transit planning. *Journal of Applied Communication Research, 40*(3), 307–328.

Chen, H., Wigand, R. T., & Nilan, M. S. (1999). Optimal experience of web activities. *Computers in Human Behavior, 15*(5), 585–608.

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety: The experience of play in work and games*. San Francisco: Jossey-Bass.

Csikszentmihalyi, M. (1990). *Flow*. New York: Harper and Row.

Csikszentmihalyi, M., & Csikszentmihalyi, I. (1988). *Optimal experience*. Cambridge: Cambridge University Press.

Dimoka, A., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Gefen, D., Gupta, A., Ischebeck, A., Kenning, P., Pavlou, P. A., Müller-Putz, G., Riedl, R., vom Brocke, J., & Weber, B. (2012). On the use of neurophysiological tools in IS research: Developing a research agenda for NeuroIS. *MISQ, 36*(3), 679–702.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM, 54*(4), 86–96.

Eger, N., Ball, L. J., Stevens, R., & Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through eye-movement replay. In *Proceedings of the 21st British HCI group annual conference on people and computers: HCI… but not as we know it,* British Computer Society, 1, pp. 129–137.

Eisenberger, R., Jones, J. R., Stinglhamber, F., Shanock, L., & Randall, A. T. (2005). Flow experiences at work: For high need achievers alone? *Journal of Organizational Behavior, 26*(7), 755–775.

Engeser, S., & Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion, 32*(3), 158–172.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Finneran, C. M., & Zhang, P. (2005). Flow in computer-mediated environments: Promises and challenges. *Communications of the Association for Information Systems, 15,* 82–101.

Guo, Y. M., & Poole, M. S. (2009). Antecedents of flow in online shopping: A test of alternative models. *Information Systems Journal, 19*(4), 369–390.

Hoffman, D. L., & Novak, T. P. (2009). Flow online: Lessons learned and future prospects. *Journal of Interactive Marketing, 23*(1), 23–34.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: OUP.

Howe, J. (2006). The rise of crowdsourcing. *Wired*. http://www.wired.com/wired/archive/14.06/crowds.html

Jackson, S. A., Thomas, P. R., Marsh, H. W., & Smethurst, C. J. (2001). Relationships between flow, self-concept, psychological skills, and performance. *Journal of Applied Sport Psychology, 13*(2), 129–153.

Kaufmann, N., Schulze, T., & Veit, D. (2011). More than fun and money: Worker motivation in crowdsourcing–a study on mechanical turk. In *Proceedings of the seventeenth Americas conference on information systems,* Detroit.

Keller, J., & Bless, H. (2008). Flow and regulatory compatibility: An experimental approach to the flow model of intrinsic motivation. *Personality and Social Psychology Bulletin, 34*(2), 196–209.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218.

Lakhani, K. R., & Wolf, R. G. (2005). Why hackers do what they do: Understanding motivation and effort in free/open source software projects. In J. Feller, B. Fitzgerald, S. A. Hissam, & K. R. Lakhani (Eds.), *Perspectives on free and open source software* (pp. 3–22). Cambridge, MA: MIT Press.

Malone, T. W., Laubacher, R., & Dellarocas, C. N. (2009). *Harnessing crowds: Mapping the genome of collective intelligence*. MIT Sloan research paper no. 4732-09. http://ssrn.com/abstract=1381502, http://dx.doi.org/10.2139/ssrn.1381502

Mannell, R. C., & Bradley, W. (1986). Does greater freedom always lead to greater leisure? Testing a person × environment model of freedom and leisure. *Journal of Leisure Research, 18,* 215–230.

Martin, A. J., & Jackson, S. A. (2008). Brief approaches to assessing task absorption and enhanced subjective experience: Examining "short" and "core" flow in diverse performance domains. *Motivation and Emotion, 32*(3), 141–157.

Moneta, G. B. (2012). On the measurement and conceptualization of flow. In S. Engeser (Ed.), *Advances in flow research* (pp. 23–50). New York: Springer.

Murphy, C. A., Coover, D., & Owen, S. V. (1989). Development and validation of the computer self-efficacy scale. *Educational and Psychological Measurement, 49,* 893–899.

Nakamura, J., & Csikszentmihalyi, M. (2002). The concept of flow. In C. R. Snyder & S. J. Lopez (Eds.), *Handbook of positive psychology* (pp. 89–105). New York: Oxford University Press.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54–67.

Schaufeli, W. B., Salanova, M., González-Romá, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness Studies, 3*(1), 71–92.

Sherry, J. L. (2004). Flow and media enjoyment. *Communication Theory, 14*(4), 328–347.

Wagner, C., & Prasarnphanich, P. (2007). Innovating collaborative content creation: The role of altruism and wiki technology. In *40th annual Hawaii international conference on system sciences*, pp. 18–28.

Zheng, H., Li, D., & Hou, W. (2011). Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce, 15*(4), 57–88.

Chapter 13
Modeling Dynamic Organizational Network Structure

Seokwoo Song and Seong-Hoon Choi

Abstract The organizational social networks, where the creation and recombination of knowledge typically takes place, are recognized as a crucial enabler for improving the organizational innovation and performance. While the recent research endeavors have been insightful in explaining the effect of the organizational social networks, we may need more effective tools to investigate the dynamics of the evolving network structures within an organization. Agent-based modeling has been considered a powerful tool for thoroughly studying the dynamics of the system. In this study, we propose an agent-based simulation model to provide a deeper understanding the dynamics of organizational network structures along with its task environment.

Keywords Agent-based modeling • Social network • Simulation

13.1 Introduction

In today's knowledge intensive environment, the conscious efforts of organizations to create and share knowledge among the members may have become a crucial component to sustain their competitive advantage. The creation and recombination of knowledge typically takes place within the organizational social networks that are recognized as an important factor in influencing the organizational performance (Song et al. 2007). Many researchers have studied organizational social networks in

S. Song (✉)
Department of Business Administration, Goddard School of Business & Economics,
Weber State University, 3802 University Circle, Ogden, UT 84408-3802, USA
e-mail: seokwoosong@weber.edu

S.-H. Choi
Department of Management Engineering, Sangmyung University, Seoul, South Korea
e-mail: shchoi@smu.ac.kr

the various ranges of organizational phenomena such as the impact of network diversity and density on organizational performance (Reagans and Zuckerman 2001), inter-unit resource exchange and product innovation (Tsai and Ghoshal 1998), relationship between group social structure and group effectiveness (Oh et al. 2004), network effects on knowledge transfer (Reagans and McEvily 2003), relationship between job performance and advice network structure (Sparrowe et al. 2001), and the relationship between social network ties and interpersonal citizenship behaviors (Bowler and Brass 2006).

In general, social networks within an organization are dynamically developed over time. In order to thoroughly study the effects of evolving organizational networks, longitudinal data collection may be required. Most recent studies, however, have been bounded to the cross-sectional approach (Chen et al. 2012) since labor-intensive and time-consuming efforts are required to collect consistent longitudinal data. Much attention has still been directed to exploit the organizational network approach (Kilduff and Brass 2010). While these research endeavors have been insightful in explaining the effect of the organizational social networks, we may need more effective tools to investigate the dynamics of the evolving network structure within an organization.

The recent development of mathematical and computational models, including agent-based models, has shown potential to explore the impacts and patterns of the evolving organizational social networks. Ashworth and Carley (2006) proposed some network measures and used simulation to investigate the relative effectiveness of social network theory as predictors of individual contributions to team performance. Hirshman et al. (2011) used an agent-based simulation model to implement homophily and explain the tiering behaviors in human networks. Lazar and Friedman (2007) presented an agent-based model to examine the effect of communication network structure on system-level performance. Zaffar et al. (2011) utilized an agent-based computational economics model to examine social and economic factors on the diffusion dynamics of open source software.

The goal of this study is to propose a novel agent-based modeling framework to incorporate the dynamics of organizational social networks. In addition, a typical member in an organization spends most of his/her work days interacting with other members within the same team. In the proposed model, we attempt to address these social aspects associated with team compositions within an organization. Thus, the proposed model enables us to investigate complex communication patterns within an organization, as well as to explore the social interactions in the task environments, which may lead to the organizational innovative outcomes (Lazer and Friedman 2007; Miller et al. 2006).

The remainder of this paper is organized as follows. We will first present an overview of prior studies on the agent-based model. This is followed by the design and development of the proposed agent-based simulation model, including the initialization and evolving processes. In the subsequent sections, the simulation results will be presented. Finally, the conclusions and possible future studies are discussed.

13.2 Dynamic Organizational Social Network

In this section, we propose an agent-based simulation model to incorporate the organizational task environment and organizational learning process. Agent-based modeling has been used in diverse disciplines such as economics and sociology. The framework is appropriate for modeling complex interactions which are intrinsic to a real world setting.

13.2.1 Background Literature

Song et al. (2007) argued that the organizational network structures are the basis of social capital, which allows individuals or groups to control information and access diverse and complementary resources and skills within the organization. The individuals and groups within an organization, as actors in the organizational social networks, are involved in exchanging its limited resources such as information and knowledge.

The agent-based models are recognized as ideal tools for conceptually experimenting real-world actors (Lazer and Friedman 2007). Many studies have addressed that computer simulation can be a powerful research tool for complex systems like organizations (Anderson 1999; Davis et al. 2007; Harrison et al. 2007). The agent-based approach has been used for modeling the behaviors of adaptive decision makers through interaction (Fioretti 2013; Harrison et al. 2007). Social interaction and knowledge exchange in the multi-agent systems depend upon the connection between agents. By using computational experiments, the agent-based models enable us to understand how a complex system behaves (Fioretti 2013). Further, Burk et al. (2007) proposed an actor-based model, a sort of agent-based modeling, to examine the co-evolution of social networks and individual behaviors.

There have been a number of recent empirical papers pertaining to the various applications of agent-based simulations such as knowledge exchange (Wang et al. 2009), social activity generation and scheduling (Ronald et al. 2012), computer-mediated communication (Canessa and Riolo 2006), innovation diffusion within market networks (Bohlmann et al. 2010), and innovation networks (Ahrweiler et al. 2011). Table 13.1 provides examples of these research studies.

13.2.2 Model Design

The proposed model in this study is implemented by using Arena (Kelton et al. 2001) with MS Excel for controlling input and output (see Fig. 13.1).

Table 13.1 Research on agent-based model

	Applications (tool)	Findings
Ahrweiler et al. (2011)	Innovation networks (SKIN)	Actors in knowledge intensive industries are able to compensate for structural limitations through strategic collaborations
Bohlmann et al. (2010)	Innovation diffusion within market networks	The ability to speed innovation diffusion varies significantly according to within- and cross-segment communications within a heterogeneous network structure
Canessa and Riolo (2006)	Computer-mediated communication	The CMC results generated by an agent-based model can lead to a deeper understanding of the behavior of a complex adaptive system
Hirshman et al. (2011)	Knowledge Homophily (CONSTRUCT)	Supplementing homophily with highly valued personal facts can more successfully lead to the tiering behavior
Lazer and Friedman (2007)	Network structure on system-level performance	An efficient network positively affects information diffusion, but negatively affects information diversity, which is also positively related to performance
Ronald et al. (2012)	Social activity generation and scheduling (Python)	The proposed model is most sensitive to the pair attributes of the network, rather than the global or personal attributes
Wang et al. (2009)	Knowledge sharing (Repast)	Knowledge sharing results from complex interaction between employee behavior and organizational interventions
Zaffar et al. (2011)	Open Source Software (ACE)	The controllable factors, such as interoperability costs, are major determinants of open source software diffusion

13.2.2.1 Initialization

To initialize the proposed model, we set up the initial parameters such as time lag = 100, the number of agents = 100, the number of teams = 10, and so on. We also assign some attribute values (e.g., tenure with normal distribution, $N(5, 0.5)$) to each agent. Then, to develop the relations between each agent, we follow the next steps:

1. Assign a minimum number of agents (e.g., normal distribution $N(3, 0.5)$) to each team
2. Compute the Euclidean distance, d_{ij}, between the unassigned agent i and team j.
   $d_{ij} = \sqrt{\sum_{\forall k} \left(a_{ik} - T_{jk} \right)^2}$ where a_{ik} refers to k th attribute for agent i and T_{jk} refers to k th attribute for team j
3. Assign the agent i to the team j with the smallest d_{ij}
4. Re-calculate T_{jk}
5. Repeat the previous steps (1–4) *until every agent is assigned to a single team*

In an organization, simple tasks like day-to-day operations may not require diverse and professional information or knowledge, since team members can follow

Fig. 13.1 Simulation process
(N: time lag)
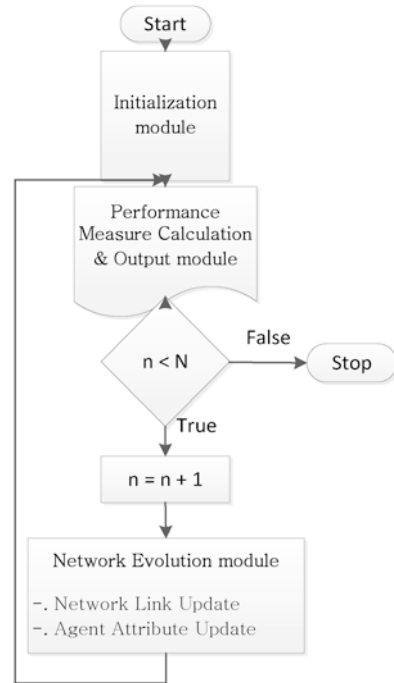
routine procedures and rules to perform the tasks. With complex tasks, on the other hand, team members may often encounter unexpected problem solving situations. Thus, a high degree of task complexity results in ambiguity and difficulty that requires new knowledge or novel solutions, and more complex tasks request for more cooperation and coordination between team members (Akgün et al. 2005). When a task involves the transmission of complex knowledge, strong ties or accurate cognitive networks between team members will prove helpful. Task complexity refers to the extent to which an actor makes an effort to solve a problem (Campbell 1988).

Once completing the assignment, we incorporate the concept of task complexity with the proposed model. We assign the value of task complexity to each agent as follows:

1. Assign a value (low, medium, or high) of task complexity *randomly* to each team
2. Limit the upper bound for each value; for example, low $= U(0,0.4)$, medium $= U(0.4,0.7)$, and high $= U(0.7,1)$
3. Determine a distribution to each value; for example, medium with $U(0.4,0.7)$ in normal distribution, $N(0.55, 0.05)$
4. Allocate the value of task complexity to each agent, depending upon the value of the team that each agent belongs to.

Figure 13.2 shows the VBA code of how we implement the above initial process.
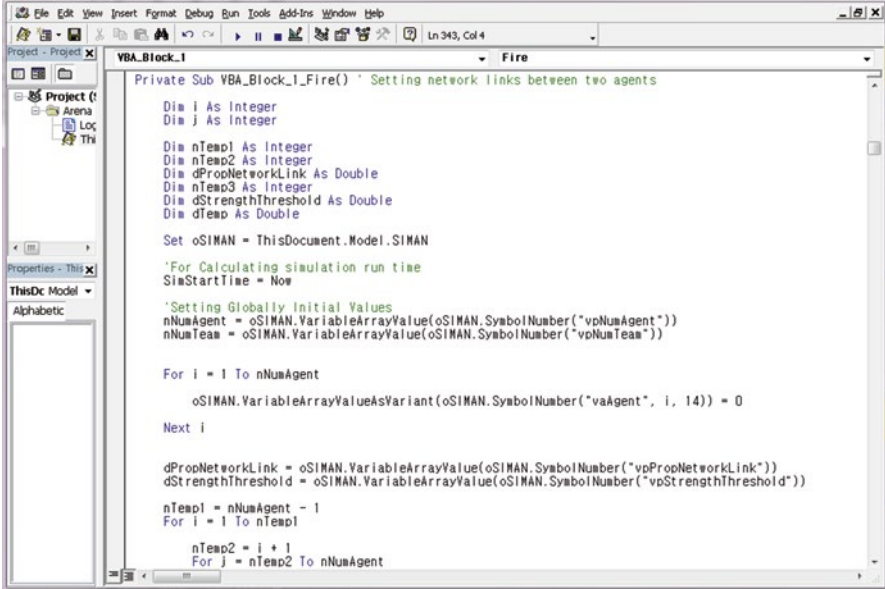
Fig. 13.2 VBA code for initial module

13.2.2.2 Evolving Networks Over Time

Jin et al. (2001) proposed the model of growing social networks, where the networks were developed depending upon the number of mutual friends and acquaintances between individuals. We adapt their idea of friendships in the proposed model. We also consider various task-related factors such as task complexity, tenure, learning, and so on.

In the proposed model, the network connections are developed at each time, computing the probability of meeting agent i with agent j. The probability, p_{ij}, depends upon the friendships and task environment of each agent, which is adapted from Jin et al. (2001).

The probability, p_{ij}, is presented as,

$$p_{ij} = S_{ij} * T_{ij}$$

$S_{ij} = f(x_i)f(x_j)g(m_{ij})$ where $f(x_i) = \dfrac{1}{e^{\beta * k(x_i)} + 1}$ and $g(m_{ij}) = 1 - (1 - p_0)e^{-\alpha m_{ij}}$ (Jin et al. 2001)

$$T_{ij} = f(c_i)f(c_j)$$

S_{ij} represents the probability of meeting two agents when each agent, x_i, already has a certain number of friends, $f(x_i)$, as well as when they share the number of mutual friends, m_{ij}, where p_0 represents the probability of meeting two agents with no mutual acquaintances (Jin et al. 2001). T_{ij} represents the probability of meeting two agents when they have similar values of task complexity. The proposed model
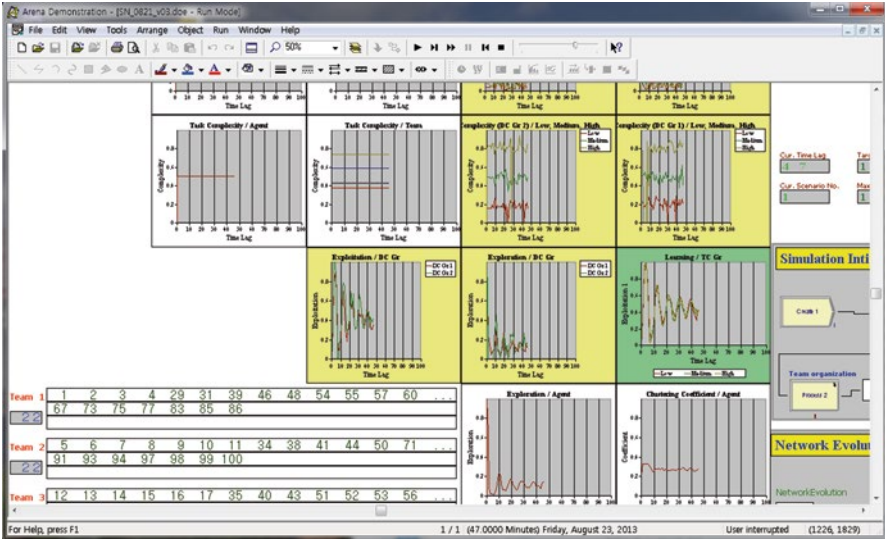
Fig. 13.3 Interface of simulation process

expects that, when two agents have similar or same values of task complexity, the chance of meeting each other may get bigger than otherwise. Thus, we consider the difference of the value of task complexity between agent i and agent j.

How do two agents meet? First, the probability of meeting two agents increases when either agent maintains a small number of friends. However, an agent may spend its time and effort to maintain the relationship with a certain number of friends. Thus, there is a limit for an agent to maintain the number of friends. The agents whose number of friends is closer to the maximum limit of the number of friends are less likely to make a new friend. Further, the probability of meeting two agents when they know each other is higher than otherwise. And, if two agents do not meet during a certain period, the relationship will be discontinued.

In addition, the probability is getting higher if they have similar degrees of task complexity. The probability of meeting two agents that belong to the same team is also higher than any agent in the different teams. Figure 13.3 shows the interface of the actual simulation process.

Since the aim of the study was to propose a dynamic model in order to elucidate the behaviors of organizational social networks, we consider a couple of network-related measures. First, we compute the clustering coefficient, which was used for small-world networks by Watts and Strogatz (1998). The clustering coefficient is the measure of the extent to which one's friends are also friends of each other (Watts and Strogatz 1998). The clustering coefficient, C_i, is presented as,

$$C_i = \frac{E_i}{\frac{k_i(k_i-1)}{2}}$$

where k_i represents the number of friends connecting with agent i and E_i represents the number of links between k_i agents.

In today's complex environment, organizations should not only continuously search for and identify new opportunities, but be able to develop its existent resources. Recent research has shown the importance of the effect of exploiting existing knowledge and/or exploring new knowledge on organizational performance and innovation (Miller et al. 2006; Raisch et al. 2009; Nielsen 2010). With the proposed model, we compute *network scope*, which corresponds to the concept of *exploration*. The *network scope* for each agent i is computed as,

$$Network\ Scope_i = \frac{new\ links_i}{total\ links_i}$$

13.2.2.3   Simulation Results

Figure 13.4 shows the trend of clustering coefficients along with the change of the number of agents (Fig. 13.4a), as well as the change of the probability, p_0, of meeting two agents when they have no previous acquaintance (Fig. 13.4b, c). For Fig. 13.4a, we use time lag = 100, $p_0 = 0.1$, $\alpha = 0.1$, and $\beta = 2$. To figure out the effect of the number of agents, we conducted the simulation changing the number of agents from 40 to 240 (increased by 40). The result showed that the values of clustering coefficients are decreasing when the number of agents increases.

Figure 13.4b, c present the trends of degree centrality as well as clustering coefficient with the values of p_0 from 0.01 to 0.30 (increased by 0.05; the number of agents = 100 and the number of teams = 10). When the values of p_0 are increasing and the values of $g(m_{ij})$ are getting bigger, the chances (i.e., p_{ij}) that two agents meet are also likely to be increasing. In Fig. 13.4, when the values of p_0 are increasing, the values of clustering coefficient are decreasing (Fig. 13.4c), while the values of degree centrality are increasing (Fig. 13.4b).

In addition, we investigated the trend of *network scope*. To conduct this test, we first computed *degree centrality* for each agent. *Degree centrality* refers to an agent's power or status in a network (Freeman 1979). *Degree centrality* indicates the extent to which an agent serves as a knowledge source or advisor to other agents. Then, we computed the average score of individual degree centrality, and made two groups: High degree (+1 Standard Deviation) and Low degree (−1 Standard Deviation).

Figure 13.5 shows the trend of *network scope*. The group with high *degree centrality* has higher values of *network scope* than that with low *degree centrality*. While changing the values of p_0 (from 0.05 to 0.30), we computed the values of *network scope*; High degree centrality group (average *network scope* = 0.152–0.173) and Low degree centrality group (average *network scope* = 0.077–0.086). Further, we computed the number of agents for each group who engage in the different level of task complexity (TC). We found: High degree centrality group (low TC = 2.67, medium TC = 8.17, high TC = 3.00) and Low degree centrality group (low TC = 3.50, medium TC = 5.00, high TC = 2.17).
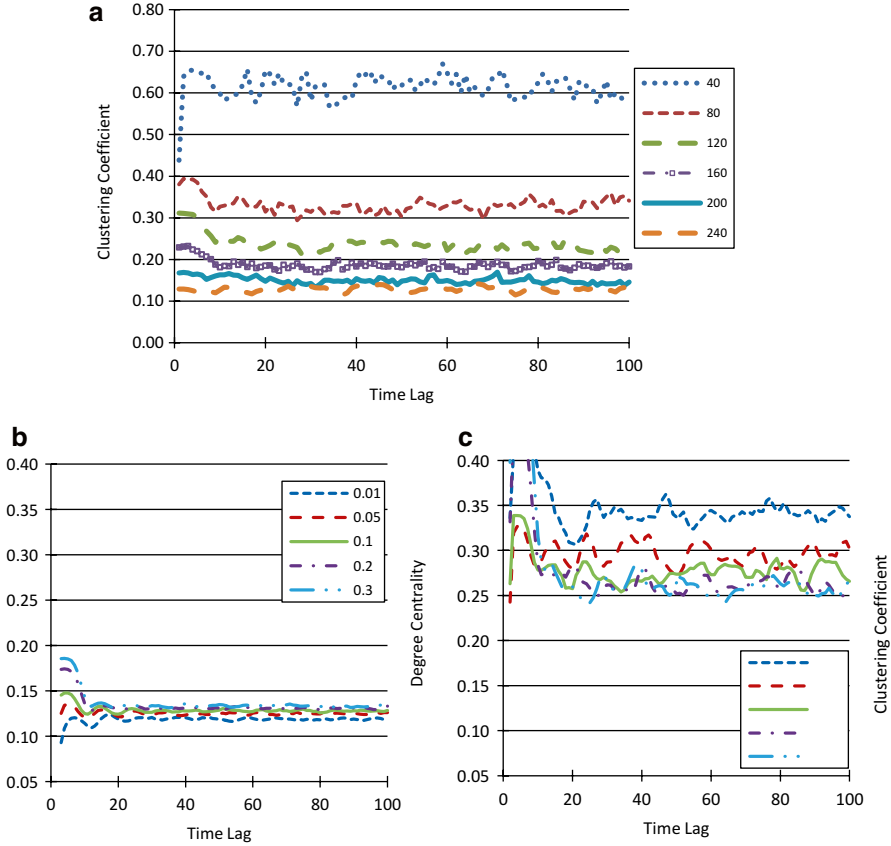
Fig. 13.4 Simulation process. (**a**) Changing the number of agents (**b**) Changing the value of p_0 (**c**) Changing the value of p_0

13.3 Discussion

The computational examples described in the previous section are the parts of various outcomes generated by the proposed model. As expected, our findings (see Fig. 13.4a) implied that the bigger the size of an organization, the fewer social interactions between members.

We also found that, when the chances of meeting two members that have no mutual acquaintance are increasing, clustering coefficient are becoming smaller (see Fig. 13.4c), while degree centrality are getting larger (see Fig. 13.4b), akin to the prior study by Lin (2012) showing the negative correlation between degree centrality and cluster coefficient. Higher clustering coefficients are more likely to lead to increasing happiness in the social relationships (Bliss et al. 2012) and strong bondage between members (Watts and Strogatz 1998) within an organization.
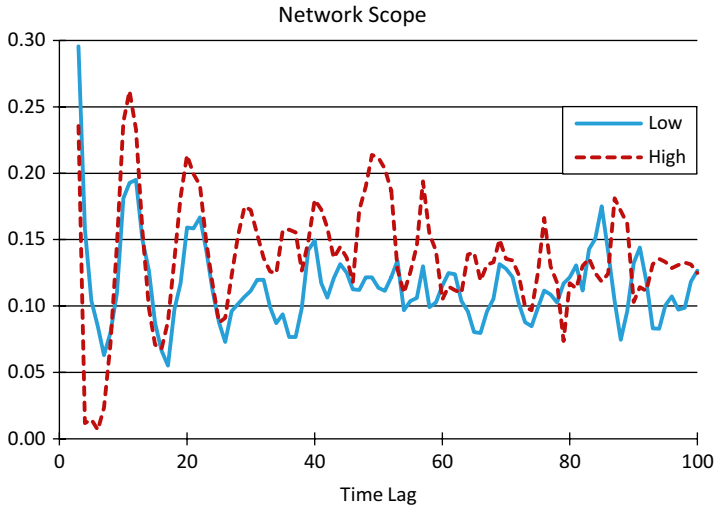
Fig. 13.5 Network scope: group by degree centrality

Members who hold higher degree centrality, on the other hand, may have positions of prominence, which enables them to maintain high communication and knowledge exchange and to have great potential to create new linkage (Song et al. 2007). Our findings implied that an organization may engage in developing different organizational cultures, depending upon its strategic direction.

Our findings of *network scope* (see Fig. 13.5) showed that the members who hold positions of prominence may be more likely to meet new members. In addition, we found that the members who hold positions of prominence involved in higher task complexity, which is consistent with the prior studies (Akgün et al. 2005) that more complex tasks request for more cooperation and coordination between team members. Our findings implied that an organization may keep promoting informal activities to maintain more social interactions between members, when the organization involves in many tasks with high task complexity.

13.4 Conclusions

Despite the growing popularity of utilizing the social network approach to examine organizational communication patterns (Kilduff and Brass 2010), we have little knowledge of the dynamics of organizational social networks. Our research attempted to provide a deeper understanding of organizational network structures along with its task environment by using the agent-based model.

Agent-based modeling has been considered a powerful tool for computational work (Epstein 1999), even with the limitations such that the model may be difficult

to validate (Canessa and Riolo 2006). Researchers addressed that agent-based modeling enables us to thoroughly study the dynamics of the system (Bonabeau 2002; Canessa and Riolo 2006). Deeper insights may be gained by considering both the task environment of the agents and the agent's network position in lateral linkages with other organizational teams.

Although the proposed model does not include performance measures, many studies showed that there are positive relationships between organizational network structure and innovative outcomes (Ahuja 2000; Song et al. 2007). Thus, our proposed model can be extended to investigate the generalizability of the influence patterns of network positions, either within or between organizations. However, there is a need to further examine the finer texture of the relationship in an intra-organizational context.

Our study has revealed some interesting patterns of *network scope*, akin to *exploration* in the previous research (Miller et al. 2006; Raisch et al. 2009; Nielsen 2010), which relates to the organizational innovative performance. These findings provide a good foundation for further inquiry. Our research is a good preliminary effort at unraveling the recursive process, suggesting a way to examine the patterns of dynamically evolving network structures within an organization.

Biography

Seokwoo Song   is Professor in Information Systems & Technologies at Weber State University. He holds a B.B.A. from Seoul National University, Korea, a MBA from Syracuse University, and a Ph.D. in Information Systems from the University of Wisconsin at Milwaukee. He has published papers in journals such as *DATA BASE for Advances in Information Systems, CACM, Information Systems Frontiers, and Journal of Knowledge Management.*

Seong-Hoon Choi   is Professor in Management Engineering at Sangmyung University, Korea. He holds a Ph.D. in Korea Advanced Institute of Science & Technology, Korea.

References

Ahrweiler, P., Gilbert, N., & Pyka, A. (2011). Agency and structure: A social simulation af knowledge intensive industries. *Computational and Mathematical Organization Theory, 17*, 59–76.
Akgün, A. E., Byrne, J., Keskin, H., Lynn, G. S., & Imamoglu, S. Z. (2005). Knowledge networks in new product development projects: A transactive memory perspective. *Information & Management, 42*(8), 1105–1120.
Anderson, P. (1999). Complexity theory and organization science. *Organization Science, 10*, 216–232.
Ashworth, M., & Carley, K. (2006). Who you know vs. what you know: The impact of social position and knowledge on team performance. *Journal of Mathematical Sociology, 30*, 43–75.

Ahuja, G. (2000). Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative Science Quarterly, 45*(3), 425–455.

Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., & Dodds, P. S. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computer Science, 3*(5), 388–397.

Bohlmann, J. D., Calantone, R. J., & Zhao, M. (2010). The effects of market network heterogeneity on innovation diffusion: An agent-based modeling approach. *Journal of Product Innovation Management, 27*(5), 741–760.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 7280–7287.

Bowler, M., & Brass, D. J. (2006). Relational correlates of interpersonal citizenship behavior, a social network perspective. *Journal of Applied Psychology, 91*, 70–82.

Burk, W. J., Steglich, C. E. G., & Snijders, T. A. B. (2007). Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors. *International Journal of Behavioral Development, 31*, 397–404.

Campbell, D. (1988). Task complexity: A review and analysis. *Academy of Management Journal, 13*, 40–52.

Canessa, E., & Riolo, R. L. (2006). An agent-based model of the impact of computer-mediated communication on organizational culture and performance: An example of the application of complex systems analysis tools to the study of CIS. *Journal of Information Technology, 21*, 272–283.

Chen, L., Gable, G. G., & Hu, H. (2012). Communication and organizational social networks: A simulation model. *Computational and Mathematical Organization Theory*. doi:10.1007/s10588-012-9131-0.

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing theory through simulation methods. *Academy of Management Review, 32*, 480–499.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity, 4*, 41–60.

Fioretti, G. (2013). Agent-based simulation models in organization science. *Organizational Research Methods, 16*(2), 227–242.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1*, 215–239.

Harrison, J. R., Lin, Z., Carroll, G. R., & Carley, K. M. (2007). Simulation modeling in organizational and management research. *Academy of Management Review, 32*, 1229–1245.

Hirshman, B. R., St. Charles, J., & Carley, K. M. (2011). Leaving us in tiers: Can homophily be used to generate tiering effects? *Computational and Mathematical Organization Theory, 17*(4), 318–343.

Jin, E. M., Girvan, M., & Newman, M. E. J. (2001). Structure of growing social networks. *Physical Review E, 64*, 046132.

Kelton, D., Sadowski, R., & Sadowski, D. (2001). *Simulation with Arena* (2nd ed.). New York: McGraw-Hill.

Kilduff, M., & Brass, D. J. (2010). Organizational social network research: Core ideas and key debates. *Academy of Management Annals, 4*(1), 317–357.

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly, 52*, 667–694.

Lin, J. (2012). Network analysis of China's aviation system, statistical and spatial structure. *Journal of Transport Geography, 22*, 109–117.

Miller, K. D., Zhao, M., & Calantone, R. J. (2006). Adding interpersonal learning and tacit knowledge to March's exploration-exploitation model. *Academy of Management Journal, 49*(4), 709–722.

Nielsen, B. B. (2010). Strategic fit, contractual, and procedural governance in alliances. *Journal of Business Research, 63*, 682–689.

Oh, H., Chung, M., & Labianca, G. (2004). Group social capital and group effectiveness: The role of informal socializing ties. *Academy of Management Journal, 47*, 860–875.

Raisch, S., Birkinshaw, J., Probst, G., & Tushman, M. L. (2009). Organizational ambidexterity: Balancing exploitation and exploration for sustained performance. *Organization Science, 20*(4), 685–695.

Reagans, R., & McEvily, B. (2003). Network structure and knowledge transfer: The effects of cohesion and range. *Administrative Science Quarterly, 48*, 240–267.

Reagans, R., & Zuckerman, E. W. (2001). Networks, diversity, and productivity: The social capital of corporate R&D teams. *Organization Science, 12*(4), 502–517.

Ronald, N., Dignum, V., Jonker, C., Arentze, T., & Timmermans, H. (2012). On the engineering of agent-based simulations of social activities with social networks. *Information and Software Technology, 54*(6), 625–638.

Song, S., Nerur, S., & Teng, J. (2007). An exploratory study on the roles of network structure and knowledge processing orientation in the work unit knowledge management. *Advances in Information Systems, 38*(2), 8–26.

Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. (2001). Social networks and the performance of individuals and groups. *Academy of Management Journal, 44*(2), 316–325.

Tsai, W., & Ghoshal, S. (1998). Social capital and value creation: The role of intrafirm networks. *Academy of Management Journal, 41*(4), 464–476.

Wang, J., Gwebu, K., Shanker, M., & Troutt, M. D. (2009). An application of agent-based simulation to knowledge sharing. *Decision Support Systems, 46*(2), 532–541.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440–442.

Zaffar, M. A., Kumar, R. L., & Zhao, K. (2011). Diffusion dynamics of open source software: An agent-based computational economics approach. *Decision Support Systems, 51*(3), 597–608.

Chapter 14
Teaching Analytics, Decision Support, and Business Intelligence: Challenges and Trends

Babita Gupta and Uzma Raja

Abstract Companies are increasingly embracing analytics to enhance business value. Academia is responding to this trend, with innovative curricula in DSS/BI/Analytics providing a variety of degree programs, minors, and certificate programs in online, traditional, and hybrid format. With BI field rapidly evolving, more universities are becoming interested in offering BI courses and programs. This necessitates innovations in BI pedagogy and materials that can best prepare students for the industry demands. Teaching material that incorporates real cases with real data from companies into the pedagogy provides the benefit to students to get high-level BI skills that companies need.

Keywords BI pedagogy • BI teaching material • BA/BI/DSS teaching • BI innovations • BI cases

Companies are increasingly embracing analytics to enhance business value. Academia is responding to this trend, with innovative curricula in DSS/BI/Analytics providing a variety of degree programs, minors, and certificate programs in online, traditional, and hybrid format. With BI field rapidly evolving, more universities are becoming interested in offering BI courses and programs. This necessitates innovations in BI pedagogy and materials that can best prepare students for the industry demands. Teaching material that incorporates real cases with real data from companies into the pedagogy benefits students get high-level BI skills that companies need.

Teaching Track Coordinators: Babita Gupta, Uzma Raja

B. Gupta (✉)
College of Business, California State University Monterey Bay,
100 Campus Center, Seaside, CA 93955, USA
e-mail: bgupta@csumb.edu

U. Raja
The University of Alabama, 2222 Columbia Drive, Auburn, AL 36830, USA
e-mail: uraja@cba.ua.edu

SIGDSA (now SIGDSA) Pre-ICIS Conference Teaching Track was organized primarily to serve as a forum for providing access to innovations in BI teaching material and best practices in BI pedagogy. Keeping in mind the need for "the IS field to expand its curriculum, attract more students, and become the academic leader in creating a high-skilled BI workforce" (Wixom et al. 2011), teaching track was specifically designed to attract papers that provided pragmatic, and hands-on teaching material incorporating real-world examples, assignments and hands-on software experiences, case studies, and datasets with an interdisciplinary focus. Across the business programs, there is an increasing need to integrate curriculum material that cuts across traditional functional areas. This is even more apparent in BI curriculum that analyzes data gathered from across an organization to provide answers to strategic questions. Thus BI curriculum should incorporate elements from other functional areas such as management, marketing, finance, operations, and IS. This is also useful in attracting students from across IS disciplines who are interested in BI technologies.

While we have made definite progress in moving towards BI curricula standards and developing content for teaching BI, there are still some challenges that remain an impediment to BI course and/or program offering. Based on the articles presented and discussions, some emerging issues are:

- Integrating real-world scenario in BI curriculum. Enriching BI curriculum with hands-on, practice based pedagogy using BI software would help in building BI programs that prepare highly skilled students who can "hit the ground running" in companies. This would prepare students to take advantage of numerous employment opportunities in BI.
- Distance learning is gaining momentum with universities striving to provide increased course (and often complete programs) offering in an on-line learning environment. This presents challenges for BI courses that require extensive lab work. Also, there is need to develop course material in BI that are better suited to on-line delivery model rather than traditional face-to-face learning environment.
- Developing truly interdisciplinary course material is still a challenge. While some universities are offering BI courses that are cross-listed across majors, by and large the courses are still offered in only one functional area.
- Developing validated curriculum standards that can be adopted for BI courses in undergraduate, Master of Science, and MBA IS programs.

The three papers presented here make significant contributions towards emerging issues in BI pedagogy.

The first paper is the Teradata University Network (TUN) Teaching Innovation award winning paper titled "Data Analysis of Retailer Orders to Improve Order Distribution" by Michelle L.F. Cheong and Murphy Choy. This is an example of how real world data can be incorporated into teaching cases and laboratory exercises. It also incorporates the area of inventory management and logistics that makes it relevant to the market demand. The teaching case described a data and decision analytics framework used at a leading integrated distribution and logistic services provider, IDD. Students analyze the problem faced by IDD through the Data and Decision Analytics Framework. They identify the actual cause of business problems

by collecting, preparing, and exploring data to gain business insights, before proposing what objectives and solutions can and should be implemented to improve the order distribution for a logistics service provider. It demonstrates how insights can be obtained and new solutions can be proposed by integrating data analytics with decision analytics, to reduce distribution cost for a logistics company. The laboratory exercises (accompanied with the step-by-step guide) are designed to help students collect, prepare and analyze retail ordering data and to use the data and decision analytics framework to solve the problems.

The second paper titled "An Online Graduate Certificate Credential Program at the University of Arkansas" by Paul Cronan, David Douglas, and Jell Mullins, shares experiences of the development of an online Business Analytics Graduate Certificate program. This is an extremely relevant issue for academic institutions since there is an increased demand for online courses and for analytics related skills in general. The paper describes how the curriculum incorporates technology, while bringing students at the same level of knowledge in technology, statistical analysis, database management systems, and data mining techniques. The paper also identifies the various activities, technologies and resources needed to establish an online Business Analytics program that could be beneficial for other programs looking to develop similar programs.

The third paper titled "Business Intelligence at Bharti Airtel Ltd" by Prabin Kumar Panigrahi, is a teaching case on business intelligence. It is based on an Indian telecommunication company, Bharti Tel, which is an early adopter of business intelligence solutions. The case describes the efforts of the company to leverage business intelligence and IT initiatives to increase business value. It describes how the company aligned IT and organizational strategy by adopting enterprise information systems. Data warehousing the business intelligence tools are then used on the data collected through the information systems to obtain optimum results. The case study is accompanied with the teaching note, listing the case synopsis, teaching objectives, and some suggested assignment questions.

All three of these papers add value to BI pedagogy by addressing related, yet diverse issues facing the academic community to meet the growing demands of teaching material in this area. These papers add to the growing body of work by offering insights into hands-on in class experiences and challenges and offering suggestions to facilitate faculty teaching in BI area.

As Big Data continues to explode, the need for innovation in teaching methods and tools is also growing. Events such as the SIGDSS Teaching Track allow academia, practitioners, and students to share their experiences, concerns and find real-world based solutions. TUN (http://www.teradatauniversitynetwork.com/) is an example of a free web-based portal that provides teaching and learning tools for faculty and students. Its mission is "*To be a premier academic resource for integrated data warehousing, big data analytics, and business applications. To build an international community whose members share their ideas, experiences, and resources. To develop curriculum for MIS, Marketing and CS faculty resulting in graduates who are fully prepared to tackle the world of big data and analytics. To serve as a bridge between academia and the world of practice through connections with Teradata's customers.*"
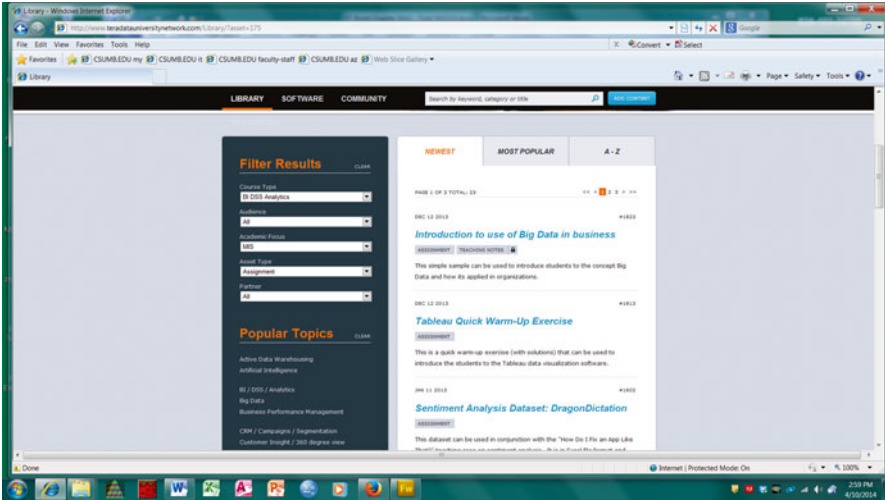
Fig. 14.1 Assignments for use in BI/DSS/Analytics courses at TUN site
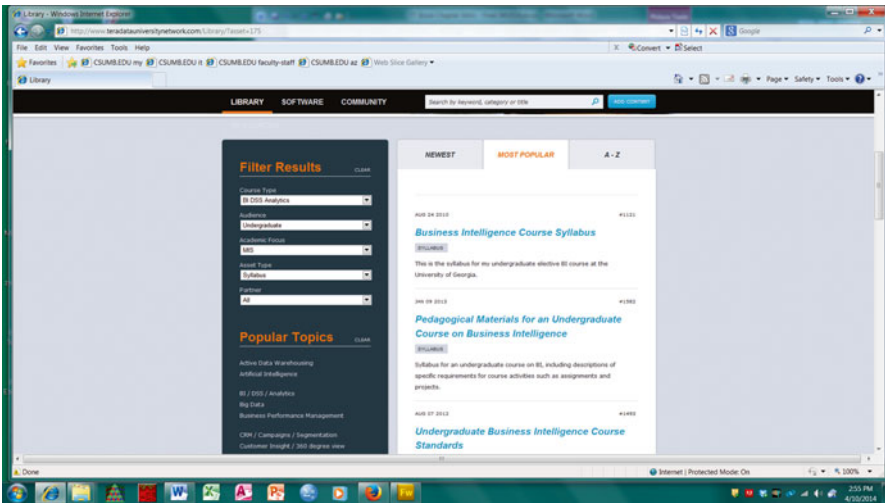

Fig. 14.2 Course syllabi being used in BI/DSS/Analytics courses at TUN site

TUN is used across the globe by various institutions. The site provides content related to BI, big data, decision support systems, data warehousing, and analytics at all levels of teaching (graduate, undergraduate, and executive level). Faculty from across the globe have built an impressive library of content that is available to all free of charge to incorporate in their courses. A faculty can search the library using various filters. For example, searching for assignments for use in BI/DSS/Analytics with MIS focus provides many results (see Fig. 14.1):

Another search for undergraduate course syllabi produces a list of resources (see Fig. 14.2):

Since TUN is an academia led initiative, it ensures that the content is relevant to the evolving teaching needs. We hope that with the three papers that follow, the academic community is further enriched with relevant teaching material to bring to the classroom.

Biography

Babita Gupta is a Professor of Information Systems at the College of Business, California State University Monterey Bay. Her research interests are in the areas of business intelligence, online security and privacy, adoption of ICTs in government, and role of culture in IT. She has published in journals such as the Communications of the Association for Information Systems (CAIS), Journal of Electronic Commerce Research (JECR), the Journal of Strategic Information Systems (JSIS), the Communications of the ACM (CACM), the Journal of Industrial Management and Data System (IMDS), and the Journal of Information Technology Cases and Applications (JITCA). She also serves on the Advisory Board of the Teradata University Network.

Uzma Raja is an Associate Professor of Management Information Systems at the Culverhouse College of Commerce, The University of Alabama. Her research interests are in the areas of Information Systems Evolution, Telemedicine, Data and Text Mining and Software Development. She has published in journals such as IEEE Transactions of Software Engineering, Decisions Science Journal, Empirical Software Engineering, Journal of Software Maintenance and Evolution and Information Resource Management Journal.

Reference

Wixom, B., Ariyachandra, T., Goul, M., Gray, P., Kulkarni, U., & Phillips-Wren, G. (2011). The current state of business intelligence in Academia. *Communications of the Association for Information Systems (CAIS), 29*(article 16), 299–312.

Chapter 15
Data Analysis of Retailer Orders to Improve Order Distribution

Michelle L.F. Cheong and Murphy Choy

Abstract Our paper attempts to improve the order distribution for a logistics service provider who accepts order from retailers for fast moving consumer goods. Due to the fluctuations in orders on a day to day basis, the logistics provider will need the maximum number of trucks to cater for the maximum order day, resulting in idle trucks on other days. By performing data analysis of the orders from the retailers, the inventory ordering policy of these retailers can be inferred and new order intervals proposed to smooth out the number of orders, so as to reduce the total number of trucks needed. An average of 20 % reduction of the total number of trips made can be achieved. Complementing the proposed order intervals, the corresponding new proposed order size is computed using moving average from historical order sizes, and shown to satisfy the retailers' capacity constraints within reasonable limits. We have successfully demonstrated how insights can be obtained and new solutions can be proposed by integrating data analytics with decision analytics, to reduce distribution cost for a logistics company.

Keywords Data analytics • Decision analytics • Order distribution • Inventory policy inference

15.1 Introduction

Third party logistics companies (3PL) are often faced with the challenges of managing the supply chain efficiency for their clients. For a 3PL who acts as the middle man for the distribution of goods for the brand owner to the retailers, several key performance indices (KPIs) are tracked as part of the service level agreement with their clients. One such KPI is the on-time delivery of orders to the retailers. Late deliveries will affect the sales of the products and may even affect market share of the product.

M.L.F. Cheong (✉) • M. Choy (✉)
School of Information Systems, Singapore Management University,
80 Stamford Road, Singapore 178902, Singapore
e-mail: michcheong@smu.edu.sg; goladin@gmail.com

Fluctuating orders from retailers made on-time delivery a
constant challenge for IDD.

| Principal Accounts (Brand owners) | ← Shipment order — IDD Distribution/ Merchandizing Division | ← Purchase order — Retail Stores |

Shipment order

Goods Receipt

Purchase order

Goods Delivery
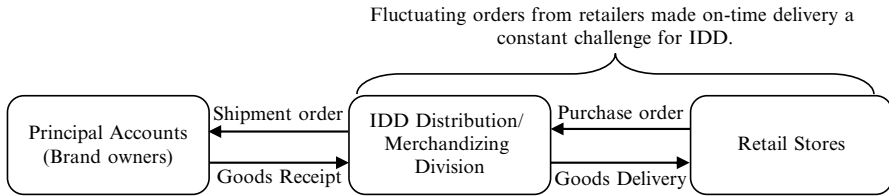
Fig. 15.1 IDD as middle-man for sales & distribution of goods

IDD is a leading integrated distribution and logistics services provider with its headquarter in Hong Kong. IDD provides a full suite of integrated distribution services covering Logistics, Distribution, Manufacturing and International Transportation. The Distribution/Merchandising division plays the middle man role (see Fig. 15.1) in distributing products for their principal accounts (brand owners) to retail stores. Products include food items such as corn flakes and chocolates, and health and beauty items such as toothpaste and shampoo.

The division was often faced with fluctuating orders from the retailers and it did not know how to best manage these fluctuations except to try its best to deliver the orders on time, and face possible penalties from the clients in case of underperforming the contracted KPI. The division wished to understand the fluctuations in orders through analysis of data captured in their IT systems. Through proper data analysis, the division hoped to gain insights on the order behavior of the retailers and propose alternative solution to achieve a win-win situation for the retailers and itself.

15.2 Literature Review

Previous work done on the fulfillment of orders from the upstream supplier or manufacturer to the downstream retailers in a two-stage supply chain under stochastic demand are often focused on sharing of Point-of-Sales (POS) information and implementing Vendor Management Inventory (VMI) so that the supplier can supply the right quantity at the right time to the retailers.

Many papers have highlighted the benefits of information sharing including reduced inventory, daily administration costs and delivery costs. Lee et al. (2000) modeled a two-stage supply chain with one manufacturer and one retailer, to quantify the benefits of information sharing and to identify the drivers that have significant impacts. They showed that manufacturer can obtain larger reductions in average inventory and average cost when the underlying demand is highly correlated over time, highly variable, or when the lead time is long. However, Raghunathan (2001) showed that sharing of demand information is of limited value when the parameters of the demand process are known to both parties, under AR(1) demand with a nonnegative autocorrelation coefficient. The reason is that the manufacturer can forecast the demand information shared by the retailer with a high degree of accuracy using retailer order history, rather than using only the most

recent order from the retailer to forecast the future orders. The accuracy increases monotonically with each subsequent time period. Consequently, the value of information shared by the retailer decreases monotonically with each time period, converging to zero in the limit. Thus, if the manufacturer uses its available information intelligently, there is no need to invest in inter-organizational systems for information sharing purposes.

Yu et al. (2002) also modeled the two-stage supply chain of a beauty product supplier and a retail store. They found that increasing information sharing will lead to Pareto improvement (at least one member in the supply chain is better off and no one is worst off) in the performance of the entire supply chain. Cheng and Wu (2005) extended the two-stage supply chain to consider multiple retailers and allowed correlation of orders to be negative, an extension from Yu et al. (2002). They introduced three different levels of information sharing from level 1 with only knowing retailers' order information; to level 2 with knowing both the retailers' order and customer demand information; and finally to level 3 with real-time information of customer demand through EDI. The optimal inventory policy under each of them was derived. Finally, they showed that both the inventory level and expected cost of the manufacturer decrease with an increase in the level of information sharing. However, they also showed that there was no difference between the inventory level and expected cost of the manufacturer for levels 2 and 3 of information sharing. This implied that there was no need for real-time sharing of demand information or VMI implementation for a two-stage supply chain.

Steckel et al. (2004) stated that whether the sharing of POS information is beneficial or not depends on the nature of the demand pattern represented by the POS information. If the demand pattern conveys continual change in ultimate downstream customer demand, the POS information can in fact distract the upstream decision maker from the more relevant information available from the orders placed by the downstream agent and the supply line. Gaur et al. (2005) extended the results of Raghunathan (2001) to cases in which demand is (AR(p), p > 1) or (ARMA(p, q), p > 1, q > 1). They found that the value of sharing demand information in a supply chain depends on the time series structure of the demand process. When both the demand process and the resulting order process are invertible, demand can be inferred by the manufacturer without requiring further information from the retailer. When demand is invertible but the resulting order process is not, sharing demand information is necessary. They proposed that the demand process is inferable from retailer's order quantity, if the upstream manufacturer's forecast of demand obtained by observing retailer's order quantity, converges almost surely to the actual realization of the demand as time t tends to infinity.

Williams and Waller (2010) compared the order forecasts for the highest echelon in a three-stage supply chain, using POS data versus using order history for cereal, canned soup and yogurt. Their results show that order forecast accuracy depends largely on the product characteristics (seasonal or not) and forecast horizon. In general, POS data produces a better forecast. However, for canned soup which is a seasonal product, POS data did not outperform order history for short term forecasting; whereas and for yogurt which is a short-life span product, POS data performs almost the same as order history.

In our case, IDD did not have any Point-of-Sales (POS) data or shared demand information from the retailers, thus IDD was unable to know or infer the actual demand. Instead, we hope to perform data analysis on historical order information to infer the inventory policies of downstream retailers, and to propose new order intervals and order sizes from historical order data to reduce distribution cost. By playing a proactive role in recommending order interval and the corresponding order size, the retailers need not place order actively, and IDD can better plan distribution to reduce cost.

We could only find two pieces of prior work which have similar objectives like ours to use data analysis to improve supply chain performance. Hausman et al. (1973) analyzed the demand data for 126 women's sportswear over 18 months to obtain three different data-generating processes, (1) ratios of successive forecasts are distributed lognormally; (2) ratios of successive forecasts are distributed as t (Student); and (3) actual demands during unequal time periods are distributed as negative binomial. They concluded that negative binomial was most closely representing the underlying process and simple to adapt to a decision model. Johnston et al. (2003) examined the order size of customers to improve the supply chain. The specific activity mentioned in the paper was that items with intermittent demand, the size of customer orders is required to produce an unbiased estimate of the demand. Also the knowledge of the distribution of demand is important for setting the maximum and minimum stock levels. Both works did not continue to use results of the analysis to make further supply chain related decisions. We think that we are the first to integrate data analytics and decision analytics, where historical data was analyzed to obtain insights to support decision making to improve the supply chain.

Our paper is organized as follow. Section 15.3 will describe the data analysis process to infer the inventory policy of the retailers. Based on the results obtained in Sect. 15.3, we propose a distribution strategy in Sect. 15.4. Based on the proposed distribution strategy in Sects. 15.4, 15.5 and 15.6 will compute the new proposed order interval and order sizes respectively. Section 15.7 aims to assess if the new proposed order sizes will violate retailers' capacity constraint. Section 15.8 compares the number of delivery trips based on the proposed strategy with historical data. Finally, Sect. 15.9 provides the conclusions.

15.3 Data Analysis of Retailer Orders to Infer Inventory Policy

The two sets of data (see Appendix) used for analysis were *Logistic data* and *Store Location data* for a cornflakes product (with each different packaging of the same product represented as a different SKU Code). *Logistic data* provided information on Retailer (identified by CustomerNo), SKU Code, SKU Description, Order Date, Order Quantity, Delivery Date, Delivery Status, Shipped Date, and Shipped Quantity; while *Store Location data* provided the Store Code (identified by Shiptocode), Store Name and Location in geo-information format. In total, there are 326 unique retailers, 191 unique SKU Codes, and 2,681 order records.

With only the historical purchase order information, the initial analysis aimed to categorize the retailers into two possible inventory policies namely, Periodic Review (PR) and Continuous Review (CR). The following assumptions were made:

1. The raw *Logistic data* was reconfigured into a new table with the number of orders for each day of the week (Monday, Tuesday, etc.) for each retailer using Order Date, regardless of the SKU item and order size.
2. Since the objective was to understand the ordering behavior of the retailers, the actual SKU item ordered is immaterial. The analysis result in the appendix supported that the ordering behavior of the retailer was independent of the SKU item ordered.
3. The order size is determined when the retailer has decided to place an order, so it is not the cause for placing order, but rather the result of placing order. Thus, when analyzing the ordering behavior, the order size was not considered. However, the order size would be computed after the order policy and order interval were determined.
4. Without loss of generality, we assumed zero delivery lead time, that is, Delivery Date is the same as Order Date. From the actual data, Delivery Date could be different from Order Date due to planned or unplanned delays.

    (a) Planned delay is usually represented by a fixed delivery lead time T days. As we are only concerned with the delivery of the orders instead of the inventory levels of IDD and the retailers, we can apply the analysis results to positive lead time T by simply shifting the results by T days.
    (b) Unplanned delay is usually due to operational inefficiencies with too many causes, and will not be included as part of the analysis.

5. Only retailers with at least ten orders were included in the analysis to ensure validity of the data analysis.

Based on the assumptions, the data were reconfigured according to day of week j. To explain the data analysis performed, we define the following notations:

- $i =$ Retailer index number, $i = 1$ to I
- $j =$ Day of week corresponding to the calendar date. $j = 1$ to 7, where $1 =$ Monday, $2 =$ Tuesday and so on. Note that there might be several orders by the same retailer i on different calendar dates which correspond to the same day of week j.
- $\bar{O}_{ij} =$ Set of orders by retailer i on day of week j
- $M_{ij} =$ Number of orders by retailer i on day of week j, where $M_{ij} = |\bar{O}_{ij}| \geq 0$
- $\bar{R}_i =$ Set of all the orders placed by retailer i.

$$\bar{R}_i = \bar{O}_{i1} \cup \bar{O}_{i2} \cup \bar{O}_{i3} \ldots \cup \bar{O}_{i7}$$

- $N_i =$ Number of orders by retailer i, where $N_i = |\bar{R}_i| \geq 0$
- $X_{ij} =$ Ratio of the number of orders placed by retailer i on day of week j and the total number of orders placed by retailer i.

$$X_{ij} = \frac{|\bar{O}_{ij}|}{|\bar{R}_i|} = \frac{M_{ij}}{N_i}$$

- Y_{iw} = Sum of any two ratios X_{ij} of retailer i for any 2 days of week j, where $w = 1$ to $^7C_2$ represents the combination index number and there are $^7C_2 = 21$ unique combinations.

  The two possible inventory policies considered are:

1. *Periodic Review (PR)* – This policy refers to reviewing the inventory level after a fixed interval period and placing the order quantity sufficient to fill up to the order-up-to level. Usually, small retailers who cannot afford the time and effort to review their inventory on a continuous basis will adopt the Periodic Review Policy. By analyzing the percentage of orders on each day of the week, we could infer the day which the retailer usually placed order.

Rule 1: Periodic Review with Single Dominant Day
If there exist a $Max_j(X_{ij}) > X_{cut}$, then retailer i is assumed to employ the periodic review policy on the dominant order day j, with a confidence interval of $(1-\alpha)\%$ and level of significance of $\alpha\%$.

In our paper, we have selected $X_{cut} = 40\%$ and state that if there exist a $Max_j(X_{ij}) > 40\%$, then retailer i is assumed to employ the periodic review policy on the dominant order day j, with more than 93.48 % confidence that the observation did not occur by chance with level of significance less than 6.52 %. Refer to Appendix for proof.

Rule 2: Periodic Review on 2 Days, But with Single Dominant Day
If there exist a $Max(Y_{iw}) > Y_{cut}$, then retailer i is assumed to employ the periodic review policy on 2 days of the week represented by the combination index w, with a confidence interval of $(1-\alpha)\%$ and level of significance of $\alpha\%$. For this combination w, if $X_{iq} > X_{ir}$ where q and r are the days of week represented by combination w, then q will be the dominant order day.

In our paper, we have selected $Y_{cut} = 60\%$ and state that if there exist a $Max(Y_{iw}) > 60\%$, then retailer i is assumed to employ the periodic review policy on 2 days of the week represented by the combination index w, with more than 97.67 % confidence that the observation did not occur by chance with level of significance less than 2.33 %. For this combination w, if $X_{iq} > X_{ir}$ where q and r are the days of week represented by combination w, then q will be the dominant order day.

2. *Continuous Review (CR)* – This policy refers to continuously reviewing the inventory level and order only when the inventory level reaches the reorder point, regardless of the day of week. Usually, larger retailers who have a warehouse and inventory management team can afford to continuously review their inventory and adopt the Continuous Review policy. Similarly, by analyzing the percentage of the total number of orders on each day of the week, we could infer that the retailers who adopted the Continuous Review policy did not have a specific day to place order, so their orders were evenly spread over 7 days.

Figure 15.2 below shows two typical retailers. The blue histogram shows a Periodic Review retailer who placed about 90 % of his orders on Monday, while the red histogram shows a Continuous Review retailer who placed orders evenly on every day of the week.
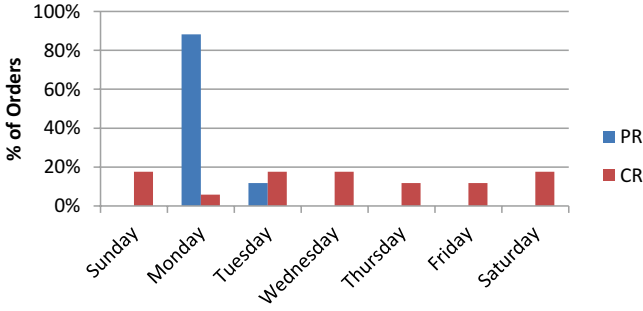
Fig. 15.2 Example of periodic review and continuous review policy retailers
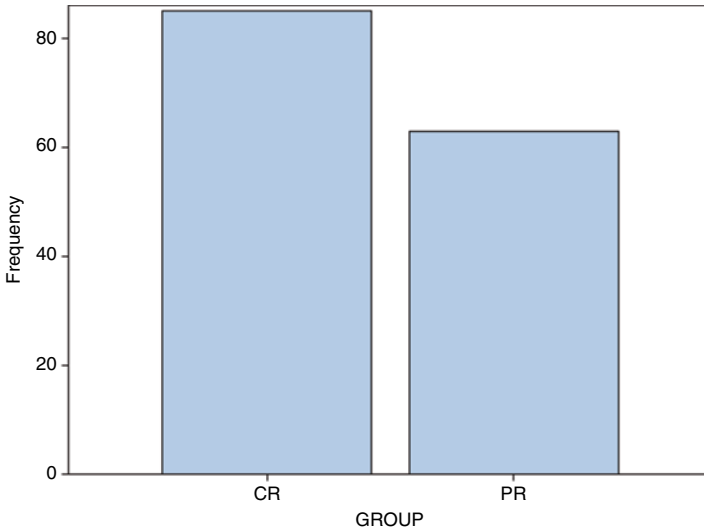
Fig. 15.3 Frequency count of retailers for different inventory policies

To compute the frequency counts for the different number of retailers for each inventory ordering policy, we adopt the following notations:

- $\bar{P}$ = Set of retailers i who employed the periodic review policy based on Rule 1 and Rule 2

$$\bar{P} = \left\{ i \mid \exists\, Max_j\left(X_{ij}\right) > 0.4\, or\, Y_{iw} > 0.6 \right\} = \bar{C}^{'}$$

- $\bar{C}$ = Set of retailers i who employed the continuous review policy

$$\bar{C} = \left\{ i \mid \vee\, Max_j\left(X_{ij}\right) > 0.4\, or\, Y_{iw} > 0.6 \right\}$$

Our result in Fig. 15.3 shows that most of the retailers employed the Continuous Review policy, that is, $|\bar{C}| > |\bar{P}|$. Since these Continuous Review policy retailers accounted for the bigger portion of the business and orders from them are rather even,

they will form the base load of orders for distribution requiring an almost fixed number of trucks, while the orders from the Periodic Review policy customers will be added on top of the base load, needing the additional trucks.

15.4 Distribution Planning Strategy

After establishing the number of retailer adopting either the Continuous Review (CR) or Periodic Review (PR) policy, we continue to understand how the orders from these retailers distribute across the different days of the week. As every retailer can place order for more than one product, we will define a retailer-product combination since we are only interested to know on which day of the week the retailers place their orders and not what products they order. Each retailer-product combination refers to a particular retailer ordering a particular product. By splitting these retailer-product combination by retailers, Fig. 15.4 shows the distribution for Continuous Review policy retailers (blue bars) which appears to be evenly spread out from Monday to Friday, while the distribution for Periodic Review policy retailers (red bars) has highs and lows from Monday to Friday. This prompted that the fluctuations in orders were caused primarily by the Periodic Review policy retailers. Such fluctuations of orders day to day, will result in needing different number of trucks for each day.

Focusing only on those retailers who adopt the Periodic Review policy, and based on their top order day, Fig. 15.5 shows that the maximum number of orders occurred on Monday, and this number was about twice that of Tuesday, the second highest order day. To ensure on time deliveries on Monday, IDD had no choice but to maintain a large fleet of trucks. However, on the other days of the week (Tuesday, Wednesday, etc.), a smaller number of trucks will be sufficient to complete all deliveries. This will result in excessive number of idle trucks on the other days of the week.
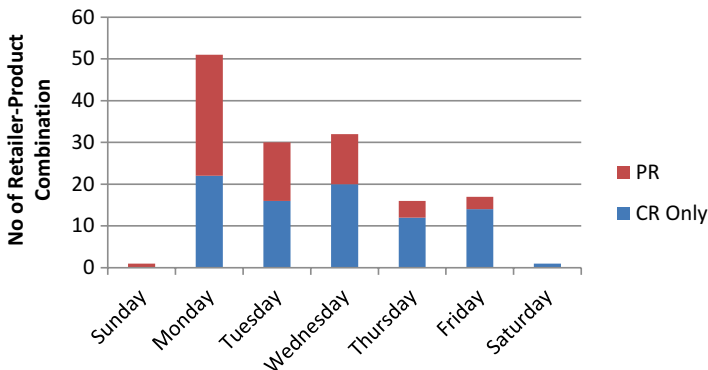
Fig. 15.4 Distribution of retailer-product combination for different day of the week
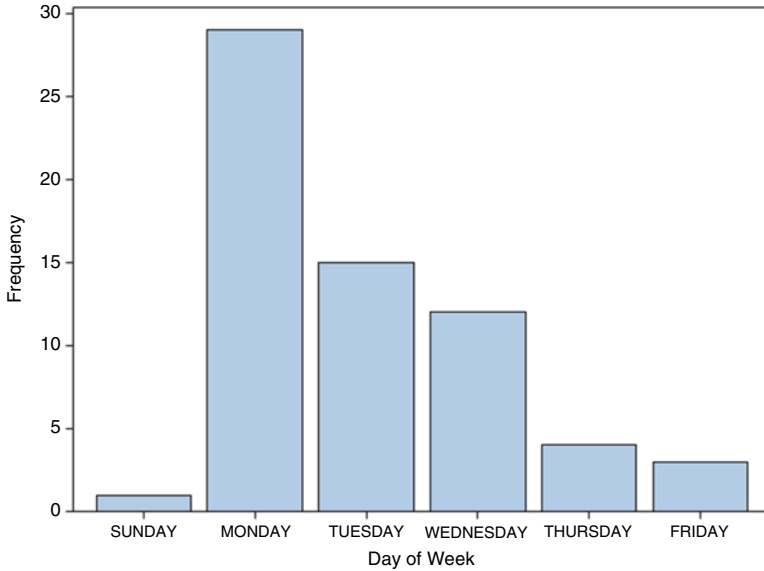
Fig. 15.5   Order frequency for each day of the week for periodic review policy retailers

For the retailers i in set $\overline{P}$,

- $\overline{P}_j$ = Set of retailers i who employed the periodic review policy on dominant day j
- $\overline{P}_1 > \overline{P}_2 > \overline{P}_3 > \overline{P}_4 > \overline{P}_5 > \overline{P}_7$. Note that there are no retailers who employed periodic review policy on Saturday.
- $\overline{P}_1 \sim 2\overline{P}_2$

IDD hoped to even out the distribution for every day of the week, so that the number of trucks used for distribution could be reduced. Since the fluctuations were caused by the Periodic Review policy retailers, the improved distribution plan would only consider smoothing out the orders from these retailers.

IDD can propose to split the retailers for Monday into two groups, each with an order interval of 14 days, instead of 7 days. Group 1 will receive goods on every 1st and 3rd Monday, while Group 2 will receive goods on every 2nd and 4th Monday. The cycle then repeats for 52 weeks in a year. For the other days of the weeks, the retailers will receive goods once a week only on their dominant day.

By carefully allocating retailers belonging to Monday into two groups, IDD can reduce the number of deliveries required for Monday, and thus reducing the total number of trucks required for the entire delivery operations. The allocation of retailers into the two groups (ideally about 50 % of Monday retailers in each group) will depend on their geographical location to minimize the travel distances. Based on the geographical location of the Monday retailers in Fig. 15.6, the Monday PR retailers are divided into five groups in (i) Kowloon, (ii) New World territory region, (iii) Yuen Long & Tuen Mun, (iv) Tung Chung, and (v) the biggest group is in the Hongkong island region. We recommend to split them into two groups,

Fig. 15.6 Geographical location of periodic review policy retailers on Monday

where the biggest group in Hongkong island will be in the first group, while the others will be in the second group, and each group will receive their orders on alternate Monday. Such a split will ensure delivery efficiency.

15.5 Implications of New Proposed Order Interval

Figure 15.7 shows the historical average order interval of Periodic Review policy retailers belonging to Monday. Note that the historical average order intervals are not in multiples of 7 days because these retailers only ordered predominantly on Mondays, but may still order on other days. Our proposed solution was to 'force' them to order only on alternate Mondays, which will make their order interval 14 days. The same principle will apply to retailers who predominantly order on other days of the week, where their average order interval will be 'forced' to be 7 days. This is known as the Power-of-Two principle where by approximating optimal order intervals to the nearest power-of-2 order interval, the total cost is guaranteed to increase not more than 6 %.

Although the total cost to the retailers will not increase by more than 6 %, there are other implications when 'forcing' them to order on alternate Mondays:

- For retailers whose historical average order interval is less than 14 days, they will be receiving orders less frequently than before, and the order size received will be larger. The main concern here would be whether the retailers would have
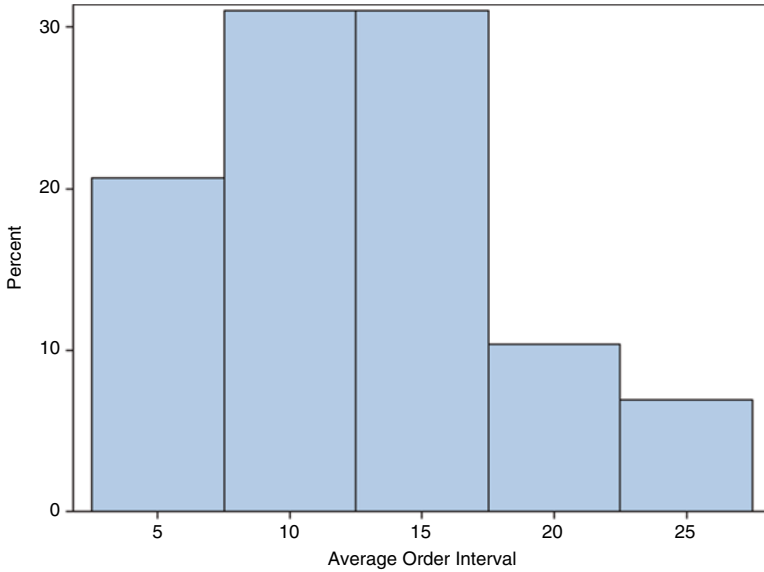
Fig. 15.7   Average order interval for periodic review policy retailers on Monday

sufficient capacity to receive the larger orders. This issue will be addressed in the next two sections.

- For retailers whose historical average order interval is more than 14 days, they will be receiving orders more frequently than before, and the order size received will be smaller. The main concern here would be whether the retailers would have the manpower to receive the orders more frequently. We will not address this issue in this paper.

15.6   Computation for Corresponding Proposed Order Size

The corresponding proposed order sizes can be computed using a moving averaging method, where the averages are computed using historical orders. Assuming historical orders in a particular period will represent future orders in the same period, the proposed order sizes are pre-computed based on historical order data for each retailer, using the proposed order interval of 7 or 14 days.

As defined previously,

- $\bar{O}_{ij}$ = Set of orders by retailer i on day of week j
- M_{ij} = Number of orders by retailer i on day of week j, where $M_{ij} = |\bar{O}_{ij}| \geq 0$
- $\bar{R}_i$ = Set of all the orders placed by retailer i
- N_i = Number of orders by retailer i, where $N_i = |\bar{R}_i| \geq 0$

So, for every retailer i in set $\overline{P}_j$, where j is the dominant order day,

- T_j = Proposed order interval.

$$T_1 = 14, \; T_2 = T_3 = T_4 = T_5 = T_6 = T_7 = 7$$

- k = Order index number of historical orders where $k = 1$ will be the first order. $k = 1$ *to* N_i
- k' = Order index number of proposed orders where $k' = 1$ will be the first order
- O_{ijk} = Order size of historical order k by retailer i for dominant order day j
- t_{ijk} = Time interval between historical order k and order k+1, by retailer i on dominant order day j. For N_i historical order, there will be $(N_i - 1)$ time intervals.
- $Q_{ijk'}$ = Proposed order size for order k' for retailer i for dominant order day j

The computation method has five main steps for any retailer i with dominant order day j, and proposed order interval T_j.

1. For initialization,

   (a) Compute the first historical average daily demand based on historical order $k = 1$

   $$Q_{ij1} = D_{ij1} * T_j$$

   (b) Compute the first proposed order size for order $k' = 1$,

   $$Q_{ij1} = D_{ij1} * T_j$$

   This proposed order size Q_{ij1} should cater adequately to demand for the first T_j days.
   (c) Let $D_{ij1} = D_{ijp}$ where the subscript p in D_{ijp} denotes previous average daily demand.

2. Compute a new average daily demand based on the closest equivalent order interval.

   $$D_{ijn} = \sum_{k=s}^{s+K} O_{ijk} \Big/ \sum_{k=s}^{s+K} t_{ijk}$$

   Where,

   - s is the starting order index number
   - K is the number of historical orders whose sum of the historical order interval matches closest the proposed time interval T_j K changes for every computation of D_{ijn}.
   - n in D_{ijn} denotes new average daily demand
   - For initialization, $s = 1$. For subsequent iterations, $s = K + 1$.

3. Compute the applied average daily demand by averaging the new average daily demand obtained in step 2, with the previous average daily demand. In case where

the actual demand is known, the actual demand for the past T_j days can replace D_{ijp} for a more accurate average demand to be applied for the next T_j days.

$$D_{ija} = \left(D_{ijp} + D_{ijn} \right)/2$$

4. Compute the adjusted proposed order size for the next T_j days

$$Q_{ija} = D_{ija} * T_j$$

By actively adjusting the proposed order size based on historical value on a moving average, the order size will be able to cater to demand changes.

5. Let $D_{ijp} = D_{ijn}$ and repeat Steps 2, 3 and 4 until the all the proposed order sizes for the entire year of 52 weeks are computed.

Example Computation Based on Table 15.1

1. Initialization

   (a) The first average daily demand was computed from the first order quantity and order interval (i.e. average daily demand = order quantity/order interval). First average daily demand = 10/5 = 2.0
   (b) Using this average daily demand, the proposed order quantity = 14 days* average daily demand = 14* 2.0 = 28. This order quantity should cater adequately to demand for the next 14 days.

2. Compute the new average daily demand based on the closest equivalent order interval. New average daily demand for the closest equivalent order interval of 15 days = (10 + 8 + 9 + 8)/(5 + 4 + 3 + 3) = 2.33
3. Compute the applied average daily demand by averaging the new average daily demand with the previous average daily demand of 2.0. The applied average daily demand = (2.33 + 2.0)/2 = 2.17
4. Adjusted order quantity for the next 14 days interval = 14* 2.17 = 30 (to nearest integer). This new order size of 30 should cater adequately to demand for the next 14 days.
5. Steps 2, 3 and 4 are repeated until the all the proposed order sizes for the entire year of 52 weeks are computed.

Table 15.1 Computation of proposed order size & adjusted order size for 14-day interval

Historical data				14 days order interval	
Order #	Order quantity	Order interval	Average daily demand	Proposed order quantity	Adjusted order quantity
1	10				
2	8	5	2.0	28	
3	9	4	2.0		
4	8	3	3.0		
5	10	3	2.7		30

15.7 Retailers' Capacity Constraint Check

Proposing a longer order interval will result in a larger order size, which may violate the storage capacities at the retail stores. However, the storage capacity at each of the retail stores was not captured in the raw data. We could however infer from the historical purchase order data, assuming that retailers who placed large order in the past would have a large storage capacity.

A measure of reasonableness will be computed as,

$$\text{Ratio } Z = \text{Maximum}\left(\text{Proposed Order Size}\right) / \text{Maximum}\left(\text{Historical Order Sizes}\right)$$

As defined previously,

- $\bar{P}$ = Set of retailers i who employed the periodic review policy
- $\bar{R}_i$ = Set of all the orders placed by retailer i
- N_i = Number of orders by retailer i, where $N_i = |\bar{R}_i| \geq 0$
- k = Order index number of historical orders for retailer i where $k = 1$ to N_i
- O_{ijk} = Order size of historical order k by retailer i for dominant order day j
- k' = Order index number of proposed orders where $k' = 1$ will be the first order
- $Q_{ijk'}$ = Proposed order size for order k' for retailer i for dominant order day j

For every retailer i in set $\bar{P}$, we determine the ratio of $Z_{k'}$ as,

$$Z_{k'} = Max_{k'}\left(Q_{ijk'}\right) / Max_k\left(O_{ijk}\right)$$

Table 15.2 shows the percentage of Periodic Review policy retailers with their respective ratio X. Ratio Group 1 has 47 % of the retailers who have Ratio $Z_{k'} < 1$, which means that the proposed order size will not exceed their storage capacity. Ratio Group 2 has 34 % of the retailers who have Ratio $Z_{k'}$ between 1 and 2, which means that the proposed order will be within 1–2 times their maximum order size, which is still reasonable. Ratio Group 3 has the remaining 19 % of the retailers who have Ratio $Z_{k'}$ above 2, which means that the proposed order size have a high chance of exceeding their storage capacity. Cost savings derived from the new distribution strategy can be passed on to these retailers to entice them to accept the new order interval and order size, especially for those in Ratio Group 3.

Table 15.2 Ratio $Z_{k'}$ of proposed order size/ maximum historical order size

Ratio group	%	Ratio $Z_{k'}$
1	47	$Z_k' <= 1$
2	34	$1 < Z_k' <= 2$
3	19	$Z_k' > 2$

15.8 Comparing Number of Delivery Trips

For retailers who employed the Continuous Review policy, there will be no change to the number of orders and thus no change to the number of delivery trips required. For retailers who employed the Periodic Review policy, the number of orders will be changed according to the proposed order intervals (14 days for Monday, and 7 days for other days of the week). The total number of delivery trips made for both policies, was compared with the original number of trips for two groups on Mondays, and 1 group each for Tuesday to Sunday, in Table 15.3.

The number of trips made based on fixed delivery day and fixed interval is reduced by about 20 % and up to 47.3 % for Sunday. The biggest improvement comes from the split of the Monday group into two groups, so that the number of trips needed on any Monday is around 1,100 trips, instead of 2,900 trips in total. This will reduce the total number of trucks required for the entire delivery operations, and in turn reduce the cost of distribution.

15.9 Conclusions

In this paper, we have demonstrated how a logistics company can make use of the data they have captured in their order system to infer the ordering behavior of their retailers. By performing data analysis to categorize the retailers into Periodic Review or Continuous Review policy groups, we could identify that the fluctuations in the number of orders were primarily caused by retailers who employed the Periodic Review policy. These Periodic Review policy retailers were then classified according to their dominant order day and the result showed that the Monday group had double the number of orders than other days of the week. The proposed solution

Table 15.3 Comparison between number of delivery trips

	Monday group 1 (14 day)	Monday group 2 (14 day)	Tuesday (7 day)	Wednesday (7 day)	Thursday (7 day)	Friday (7 day)	Saturday (7 day)	Sunday (7 day)
Original number of trips	1,433	1,455	1,469	1,271	1,074	1,007	67	165
Number of trips based on fixed delivery day and fixed interval	1,148	1,147	1,188	1,034	956	996	67	87
Reduction percentage	19.9 %	21.2 %	19.1 %	18.7 %	11.0 %	1.1 %	0 %	47.3 %

was to split the Monday retailers into two groups with order interval of 14 days, while the other retailers will have order interval of 7 days. The overall reduction in the number of trips made was about 20 % to as high as 47.3 %. The largest savings would be derived from the reduction in the number of trucks to support the entire delivery operations. We have successfully demonstrated how new solutions can be proposed by integrating data analytics with decision analytics, to reduce distribution cost for a logistics company.

15.10   Teaching Note

15.10.1   Overview

Many operations management problem ranging from demand forecasting, inventory management, distribution management, capacity planning, workforce scheduling, and queue management are usually solved using known OM/OR techniques such as algorithms, heuristics, and optimization techniques. However, such a typical OM/OR solution methodology often assumes that the actual cause of the problem is known and the problem objective is well defined.

Practitioners like us would know that real business problems do not present themselves clearly, often resulting in people solving the wrong problem. Thus, in this course, the students will be exposed to the Data and Decision Analytics Framework (Fig. 15.8) which helps the analyst to first identify the actual cause of
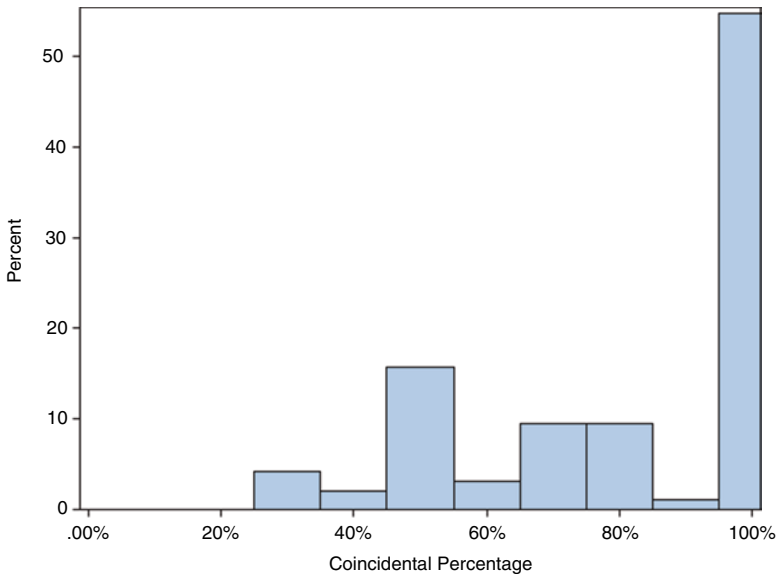
Fig. 15.8 Data & decision analytics framework

business problems by collecting, preparing, and exploring data to gain business insights, before proposing what objectives and solutions can and should be done to solve the problems.

These steps are missing in most problem solving frameworks, particularly in solving operations management problems, where the actual cause of the problem is assumed to be known and the problem objective is assumed to be well defined. However, we advocate that careful data analysis needs to be performed to identify the actual cause of business problems, before embarking on finding the solution.

15.11   Typical Flow of Classroom Activities

A typical flow of classroom activities is depicted in the flow chart in Fig. 15.9. A case usually covers multiple perspectives of operations management topics and the instructor will first cover the topics in terms of the theories and applications. When there are mathematical calculations involved, the instructor can use class activities to supplement and enhance the students' understanding.

After that, the instructor will present the case and facilitate the discussion so that the students can appreciate the case problem and think about the solution methodology according to the Data and Decision Analytics Framework. Once the students understand the intent of the case and what they are supposed to do, the instructor can facilitate the hands-on laboratory session using the step-by-step lab guide. At the end of the lab session, the instructor can instruct the students to complete assignment questions related to the case.

15.12   Introduce Operations Management Topics

For this case, the two topics to be covered include inventory management and distribution management. For inventory management, the understanding of the Periodic Review (PR) policy and Continuous Review (CR) policy should be highlighted. The instructor can ask the students the following questions to facilitate discussions:

- Give examples of goods which the periodic review policy will be more applicable
- Similarly, give examples of goods which the continuous review policy will be more applicable
- What are the advantages and disadvantages of each policy?

For distribution management, the instructor can cover the travelling salesman problem, multiple traveling salesman problem, and vehicle routing problem, introducing the different heuristics which are used to obtain good feasible solution in each problem. The main objective of distribution management is to design tours that will reduce the number of trips made when delivering goods, so as to reduce
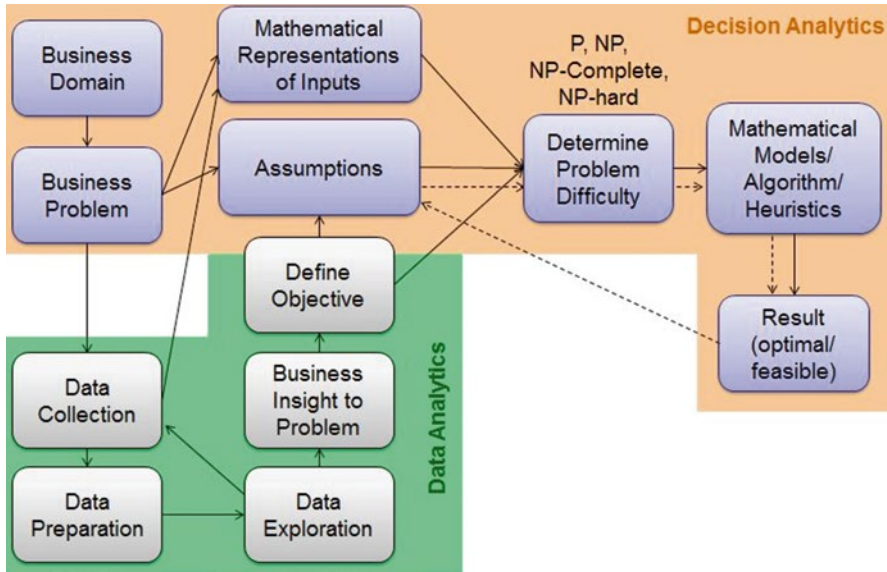
Fig. 15.9 Typical flow of classroom activities

distribution cost. The instructor can ask the students the following questions to facilitate further discussions:

- What other constraints will affect the design of the tour (time window, delivery trucks capacity constraints, client's preferences, traffic conditions)?
- What practical considerations should the vehicle routing planner consider when planning route for a particular driver (familiarity with road, ability to handle different truck size)?
- What practical considerations should the vehicle routing planner consider when planning route for a particular truck (types of goods – refrigerated or not, size of truck, maximum tonnage, door types – open at the back or at the sides)?

15.13 Conduct Case Discussion

15.13.1 Introduce the Case

The case is about IDD which is a leading integrated distribution and logistics services provider with its headquarters in Hong Kong. IDD provides a full suite of integrated distribution services covering Logistics, Distribution, Manufacturing and International Transportation.

The Distribution/Merchandising division of IDD plays the middle man role (see Fig. 15.10) in distributing products for their principal accounts (brand owners) to retail stores. Products include food items such as corn flakes and chocolates,
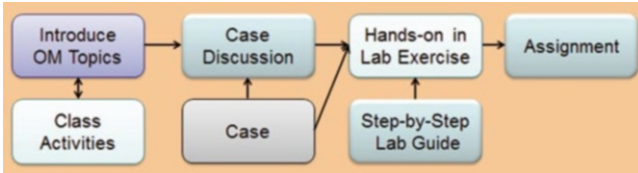
Fig. 15.10 IDD facing problem in distributing fluctuating orders to retailers
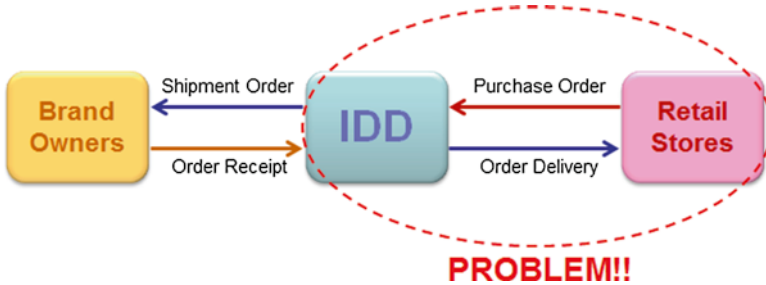
Fig. 15.11 Bullwhip effect experienced in IDD's supply chain

and health and beauty items such as toothpaste and shampoo. The division faces distribution challenges from IDD to the retailers. Orders from retailers fluctuates daily and these fluctuations resulted in the distribution team working very hard with delivery trucks rushing to deliver orders on every Monday, while on other days of the week, the team sees idle trucks parking at the warehouse un-utilized. Playing the passive middle-man role, IDD can only prepare the maximum resource capacities (e.g. drivers and trucks) in order to handle such uncertainties.

The instructor can go further to explain the bullwhip effect in supply chains which is caused by factors such as long lead time, batch ordering and demand variation. In this case, the fluctuations in the retailers orders are likely to be caused by batch ordering behavior of the retailers since demand variation on fast moving consumer goods like cornflakes and toothpaste are relatively low, as shown in Fig. 15.11.

15.13.2   *Possible Solutions and Data Provided*

After the case introduction, the instruction will ask the students to suggest possible solutions to solve the problem and for each viable suggestion, the students can discuss the pros and cons. One possible suggestion would be to implement Vendor Managed Inventory (VMI) where IDD will deliver the required quantity of products just in time, and the retailers need not place orders actively. For this suggestion, the instructor can ask the students to discuss about the pros and cons of Vendor Managed Inventory.

The Pros include:

- VMI solution will be win-win for both IDD and the retailers
- IDD can plan the deliveries better and reduce the overall cost of deliveries
- The retailers can eliminate manpower to do inventory checks and place orders

The Cons include:

- VMI implementation will require that the retailers share their Point-of-Sales (POS) data with IDD
- Due to confidentiality and trust, most retailers will not be willing to share their POS data

At this point, the instructor can highlight that IDD's IT system stores historical records of the orders from the retailers as well as the store location of each retailer provided in the Appendix of the main paper. With the order data provided (consisting data of 326 retailers and 2,681 orders), the instructor can direct the students to focus on the following four fields:

- Customer No – this is the unique customer ID
- Order Date – this is the order date
- Original Qty – this is the order quantity
- StorerClientCode – this is the store code

With the store location data provided, the instructor can direct the students to focus on the following three fields:

- Latitude – this is the latitude of the store location in geo-information format
- Longitude – this is the longitude of the store location in geo-information format
- Shiptocode – this is the store code which corresponds to StorerClientCode in the Order Data table

15.13.3   Classification Rule

After understanding the data provided, the instructor will lead the discussion on how to infer the retailers' inventory ordering behavior from using the order date. To perform the inference, the instructor needs to explain the Classification Rule (Rule 1 provided in the main paper) which is used to classify the retailers according to Continuous Review (CR) policy or Periodic Review (PR) policy.

At this point, the instructor can ask the students what if X_{cut} is chosen to be say, 60 %? Will the number of retailers categorized into PR retailers be more or fewer?

Upon using the classification rule to categorize the retailers into PR and CR policy, the instructor can explain that by plotting simple bar charts to visualize how
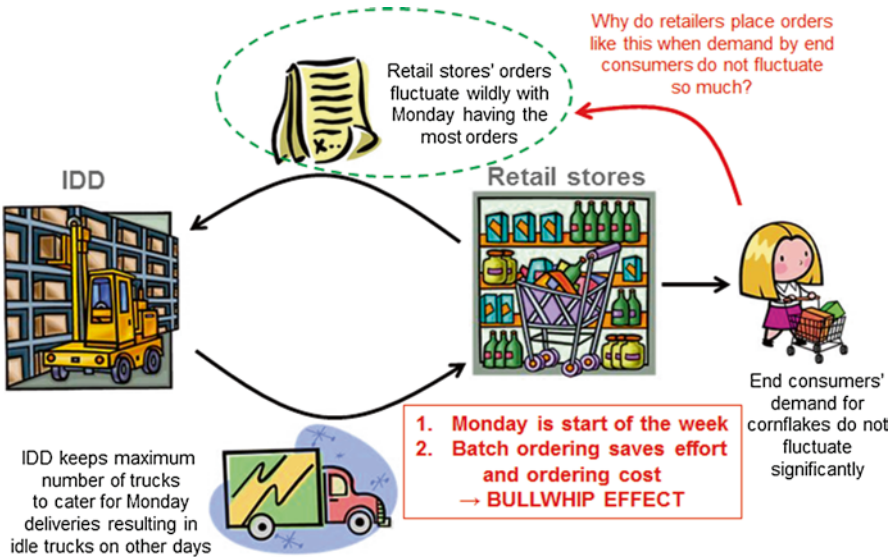
Fig. 15.12 Laboratory exercise activities

many PR and CR retailers place their orders on their dominant order day, students will be able to identify the cause of the order fluctuations and proceed to recommend a new distribution strategy.

15.14   Conduct the Laboratory Exercise

At this point, the students would have appreciated the case problem and understood that they need to perform the laboratory session with the following tasks (depicted in Fig. 15.12),

1. Infer the retailers' inventory ordering behavior by categorizing them into PR and CR according to the classification rule
2. Plot bar charts to visualize the distribution of the retailers according to their dominant order day, and use the bar charts to deduce the root cause of the order fluctuations
3. Propose new distribution strategy which can allow IDD to play a more active role to plan the delivery of the orders to the retailers on each day of the week, and propose the quantity to deliver
4. What constraints must IDD consider and how can IDD ensure that the new distribution strategy is practical (e.g. retailers' capacity challenge)?
5. Justify that the new distribution strategy will result in cost reduction.

15.15   Ensure Learning Outcomes Are Achieved

The entire case aims to achieve several learning outcomes:

1. *Exposure to supply chain business domain covering two major operations management topics including inventory management and distribution management*
   This learning outcome is achieved when the instructor covers the two operations management topics on the theories and the applications, together with class discussion and supplemented with class activities if needed.
2. *Ability to identify the actual cause of business problem by collecting, preparing, and exploring data to gain business insights, before proposing what objectives and solutions can and should be done to solve the problems using the Data and Decision Analytics Framework*
   This learning outcome is achieved when the students apply the steps in the Data & Decision Analytics Framework.
3. *Ability to propose solutions which are practical and provide cost justification*

The third learning outcome is achieved when the students perform the computations for the new proposed order size for the retailers' capacity constraint check and compute the reduction in the number of trips.

Finally, to further enhance the understanding of the case, the students can be asked to complete an assignment with the following question:

Assuming that you can dictate the type of data and information you can get from the business and you can propose a new "Order-to-Distribution-Process", propose an alternative solution to improve distribution and list the types of data needed from new the business process. Map the process flow for your proposed solution.

15.16   Appendix

1. *Logistic Data*

The logistic data contains information about the logistic transport of the goods to the retailer. Here, the retailer is identified by CustomerNo. Table 15.4 also contains some information about the expected delivery of the goods.

2. *Store Location Data*

The store location data contains the information of all the retailers' store location in geo-information format. Here in Table 15.5, the retailer is identified by Shiptocode.

3. *Proofs for Rules 1 & 2*

(a) *Proof for Rule 1: Periodic Review with Single Dominant Day*

If there exist a $Max_j(X_{ij}) > X_{cut}$, then retailer i is assumed to employ the periodic review policy on the dominant order day j, with a confidence interval of $(1-\alpha)\%$ and level of significance of $\alpha\%$.

Table 15.4 Comparison between number of delivery trips

Field name	Field description
CountryCode	Country Code
CustomerNo	Customer ID
ExpectedDeliveryDate	Expected Delivery Date of Good
OrderDate	Order Date
OrderKey	Order Key
OriginalQty	Original Order Quantity
PODDeliveryDate	Final Delivery Date
PODStatus	Final Delivery Status
PODStatusDescription	Final Delivery Status Description
PrincipalCode	Principal Code
PrincipalDescription	Principal Description
ShippedDate	Shipped Date
ShippedQty	Shipped Quantity
SkuCode	SKU Code
SkuDescription	SKU Description
StorerClientCode	Storer Code

Table 15.5 Comparison between number of delivery trips

Field name	Field description
Latitude	Latitude
Longitude	Longitude
Shiptoaddress1	Address 1
Shiptoaddress2	Address 2
Shiptocity	City
Shiptocode	Store Code
Shiptoname	Store Name
Storerkey	Storer ID
Storername	Storer Name

Consider an order from retailer i which can occur on any of the 7 days of the week.

- The probability of the order falling on a particular day of interest is $\frac{1}{7}$, and we call this the probability of success.
- Thus, the remaining probability of the order *not* falling on that particular day of interest is $\frac{6}{7}$, and we call this the probability of failure.
- This allows us to formulate a Binomial Test with $p = \frac{1}{7}$ and number of trials $= 7$, to determine the X_{cut} with the corresponding confidence interval $(1-\alpha)\%$ and level of significance $\alpha\%$.

From Table 15.6, it is observed that:

- If the percentage of occurrence of orders for a particular day of interest is 14.3 %, we are 73.65 % confident that the observation did not occur by chance with the level of significance of 26.35 %.

Table 15.6 PMF and CDF for binomial test for single day of interest

Number of occurrence on a particular day	% of occurrence	Probability Mass Function (PMF) of binomial distribution	Cumulative Distribution Function (CDF) of binomial distribution	$1 - CDF = \alpha\%$
0	0 %	0.3399	0.3399	0.6601
1	$\frac{1}{7}=14.3\%$	0.3966	0.7365	0.2635
2	$\frac{2}{7}= 28.6\%$	0.1983	0.9348	0.0652
3	$\frac{3}{7}=42.9\%$	0.0551	0.9898	0.0102
4	$\frac{4}{7}=57.1\%$	0.0092	0.9990	0.0010
5	$\frac{5}{7}=71.4\%$	0.0009	0.9999	0.0001
6	$\frac{6}{7}=85.7\%$	0.0001	1.0000	0.0000
7	$\frac{7}{7}=100\%$	Approximately 0	1.0000	0.0000

- – If the percentage of occurrence of orders for a particular day of interest is 28.6 %, we are 93.48 % confident that the observation did not occur by chance with the level of significance of 6.52 %
- – If the percentage of occurrence of orders for a particular day of interest is 42.9 %, we are 98.98% confident that the observation did not occur by chance with the level of significance of 1.02 %
- – And so on.

- In our paper, we have selected $X_{cut} = 40\%$ and state that if there exist a $Max_j(X_{ij}) > 40\%$, then retailer i is assumed to employ the periodic review policy on th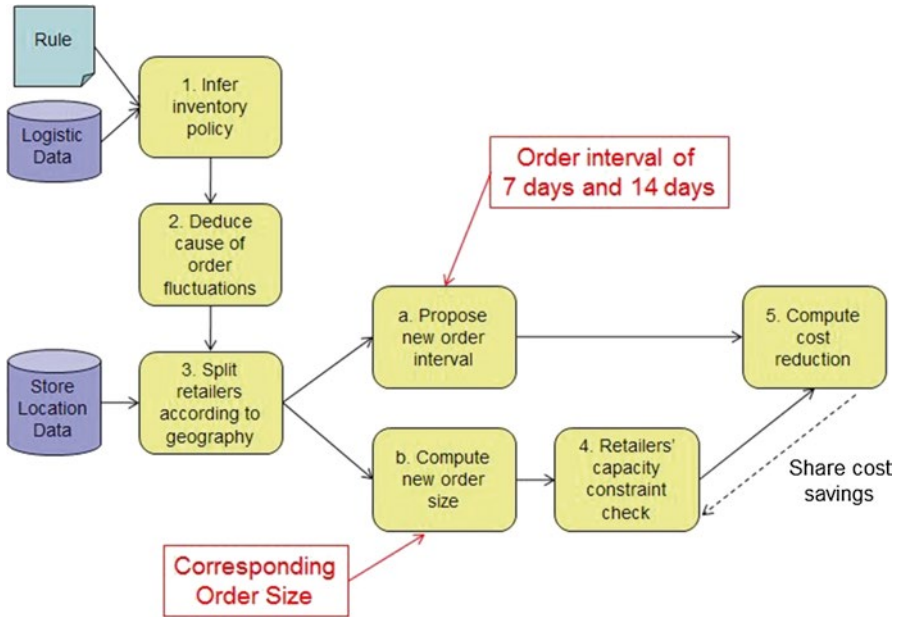e dominant order day j, with more than 93.48 % confidence that the observation did not occur by chance with level of significance less than 6.52 %.

(b) *Proof for Rule 2: Periodic Review on 2 days, but with Single Dominant Day*

If there exist a $Max(Y_{iw}) > Y_{cut}$, then retailer i is assumed to employ the periodic review policy on 2 days of the week represented by the combination index w, with a confidence interval of $(1-\alpha)\%$ and level of significance of $\alpha\%$. For this combination w, if $X_{iq} > X_{ir}$ where q and r are the days of week represented by combination w, then q will be the dominant order day.

We apply a similar Binomial Test here by grouping the 2 days of interest as 1 group, and the remaining 5 days as the other group.

Table 15.7 PMF and CDF for Binomial Test for 2 Days of Interest

Number of occurrence on a particular day	% of occurrence	Probability Mass Function (PMF) of binomial distribution	Cumulative Distribution Function (CDF) of binomial bistribution	$1 - CDF = \alpha\%$
0	0 %	0.0949	0.0949	0.9051
1	$\frac{1}{7} = 14.3\%$	0.2656	0.3605	0.6395
2	$\frac{2}{7} = 28.6\%$	0.3187	0.6792	0.3208
3	$\frac{3}{7} = 42.9\%$	0.2125	0.8917	0.1083
4	$\frac{4}{7} = 57.1\%$	0.0850	0.9767	0.0233
5	$\frac{5}{7} = 71.4\%$	0.0204	0.9971	0.0029
6	$\frac{6}{7} = 85.7\%$	0.0027	0.9998	0.0002
7	$\frac{7}{7} = 100\%$	0.0002	1.0000	0.0000

- The probability of the order falling on two particular days of interest is $\frac{2}{7}$, and we call this the probability of success.
- Thus, the remaining probability of the order not falling on that two particular days of interest is $\frac{5}{7}$, and we call this the probability of failure.
- This allows us to formulate a Binomial Test with $p = \frac{2}{7}$ and number of trials $= 7$,
- to determine the Y_{cut} with the corresponding confidence interval $(1 - \alpha)\%$ and level of significance $\alpha\%$.

From Table 15.7, it is observed that:

- If the percentage of occurrence of orders for any of the two particular days of interest is 14.3 %, we are 36.05 % confident that the observation did not occur by chance with the level of significance of 63.95 %.
- If the percentage of occurrence of orders for any of the two particular days of interest is 28.6 %, we are 67.92 % confident that the observation did not occur by chance with the level of significance of 32.08 %
- If the percentage of occurrence of orders for any of the two particular days of interest is 42.9 %, we are 89.17 % confident that the observation did not occur by chance with the level of significance of 10.83 %
- And so on.

- In our paper, we have selected $Y_{cut} = 60\%$ and state that if there exist a $Max(Y_{iw}) > 60\%$, then retailer i is assumed to employ the periodic review policy

Fig. 15.13 Coincidental analysis of dominant day for periodic review policy retailers

on 2 days of the week represented by the combination index w, with more than 97.67 % confidence that the observation did not occur by chance with level of significance less than 2.33 %. For this combination w, if $X_{iq} > X_{ir}$ where q and r are the days of week represented by combination w, then q will be the dominant order day.

4. *Coincidental Analysis of Ordering Practice for Period*

Further analysis of the Periodic Review policy retailers in the Fig. 15.13 below shows the coincidental analysis of dominant day for retailers. About 60 % of them have 100 % of their orders fixed on the same day of the week. This further justified that the ordering pattern of the Periodic Review policy retailers is independent of the SKU item ordered.

Biography

Michelle L. F. Cheong   is currently an Associate Professor of Information Systems (Practice) at the School of Information Systems (SIS) at Singapore Management University (SMU). Prior to joining SMU, she had 8 years of industry experience leading teams to develop complex IT systems which were implemented enterprise-wide covering business functions from sales to engineering, inventory management, planning, production, and distribution.

Upon obtaining her Ph.D. degree in Operations Management, she joined SMU in 2005 where she teaches the *Business Modeling with Spreadsheets* course at the undergraduate level and is the co-author of the book of the same name. She also teaches in three different master programmes at SMU on different variants of spreadsheet modeling courses covering different domains, including financial modeling, innovation modeling and IT project management. She recently designed and delivered an *Operations Focused Data Analytics* course for the Master of IT in Business (Analytics) programme at SIS. Apart from her teaching, Michelle is also the Director of Postgraduate Professional Programmes at SIS where she is in charge of two master programmes and Continuing Education & Training.

Michelle has won several awards including the *Most Promising Teacher Award* in 2007 and the *Postgraduate Professional Programme Development Award* in 2013, both from the SMU Center for Teaching Excellence. In addition, she has recently bagged the inaugural *Teradata University Network (TUN) Teaching Innovation Award* 2013, which recognizes excellence in the teaching of Business Intelligence and Business Analytics at the undergraduate, graduate and/or executive education levels.

Murphy Choy is currently a Data Analytics System and Learning Engineer at the School of Information Systems (SIS) at Singapore Management University (SMU). Prior to joining SMU, he had 4 years of industry experience in the area of risk analytics covering Asia Pacific, Middle East, Africa and Latin America. He has spearheaded several analytics initiatives in the bank and has done extensive research work for collections, recoveries and Basel II models. Murphy is especially competent in SAS software and many other analytics software, and also responsible for most of the laboratory exercises designed and taught in the Master of IT in Business (Analytics) programme.

He has served as Chairperson for the Singapore SAS user group and Section Chair for the SAS Global User group. He is pursuing his doctorate degree and his research interest is in the field of Operation Management and Text Mining. He has earned a MSc degree in Finance from University College Dublin and a BSc degree in Statistics from National University of Singapore. Murphy is also a co-author of the paper that won the inaugural *Teradata University Network (TUN) Teaching Innovation Award* 2013.

References

Cheng, T. C. E., & Wu, Y. N. (2005). The impact of information sharing in a two-level supply chain with multiple retailers. *The Journal of the Operational Research Society, 56*(10), 1159–1165.

Gaur, V., Giloni, A., & Seshadri, S. (2005). Information sharing in a supply chain under ARMA demand. *Management Science, 51*(6), 961–969.

Hausman, W. H., & Sides, R. S. G. (1973). Mail-order demands for style goods: Theory and data analysis. *Management Science, 20*(2, Application Series), 191–202.

Johnston, F. R., Boylan, J. E., & Shale, E. A. (2003). An examination of the size of orders from customers, their characterisation and the implications for inventory control of slow moving items. *The Journal of the Operational Research Society, 54*(8), 833–837.

Lee, H. L., So, K. C., & Tang, C. S. (2000). The value of information sharing in a two-level supply chain. *Management Science, 46*(5), 626–643.

Raghunathan, S. (2001). Information sharing in a supply chain: A note on its value when demand is nonstationary. *Management Science, 47*(4), 605–610.

Steckel, J. H., Gupta, S., & Banerji, A. (2004). Supply chain decision making: Will shorter cycle times and shared point-of-sale information necessarily help? *Management Science, 50*(4), 458–464, Special issue on marketing and operations management interfaces and coordination.

Williams, B. D., & Waller, M. A. (2010). Creating order forecasts: Point-of-sale or order history? *Journal of Business Logistics, 31*(2), 231–251.

Yu, C. Z., Yan, H., & Cheng, T. C. E. (2002). Modelling the benefits of information sharing-based partnerships in a two-level supply. *The Journal of the Operational Research Society, 53*(4), 436–446.

Chapter 16
An Online Graduate Certificate Credential Program at the University of Arkansas

Timothy Paul Cronan, David E. Douglas, and Jeff Mullins

Abstract Business Analytics and Big Data have become a very popular topics in recent years. Many universities are gearing up to meet the reported demand people with these skills. This paper shares background, principles, and processes in the development of an online Business Analytics Graduate Certificate Credential program consisting of four graduate courses (each three semester hours). Innovative use of technology is incorporated into all four of the courses to ensure consistency and quality content across courses. The four courses are (1) *IT Toolkit* – designed to level students (especially those students who do not have an adequate IT background), (2) *Decision Support and Analytics* – an introduction to statistical analytics with a focus on what the data is telling us, (3) *Database Management Systems* – a focus on sourcing, preparing, storing and retrieval for data and (4) *Business Intelligence* – a focus on the discovery of knowledge from data and model development using data mining including social media. Included are the efforts, activities, software, hardware, concepts, teaching philosophy, and desired outcomes for the graduate credential certificate program. The paper should be very valuable to all those teaching or planning to teach in the Business Analytics area.

Keywords Business analytics • Data • Graduate certificate

16.1 Introduction

According to a new American Management Association study (2013), demand for analytical skills are to grow sharply over the next 5 years. Many other sources confirm the impending shortage of analytical skills – perhaps driven by "Big Data." For example, McKinsey & Company (2011) state that big data is the next frontier for innovation, competition, and productivity. Academic institutions have responded to and continue to respond in many ways to this demand; from one course or two to

T.P. Cronan (✉) • D.E. Douglas • J. Mullins
Information Systems, University of Arkansas,
Business Building 204B, Fayetteville, AR 72701, USA
e-mail: cronan@uark.edu; ddouglas@walton.uark.edu; jmullins@walton.uark.edu

full-fledged degrees. Included herein is a response to this analytics demand by the Sam M. Walton College of Business at the University of Arkansas.

The initial response was creation, in 2009, of a SAS endorsed certificate program at both the undergraduate and graduate level. The certificate requirements included five components: a course in statistics, a course in programming, a course in databases/data warehouses, a course in data mining using SAS Enterprise Miner, and that the student must have worked on a project that spanned a semester. The continuing demand for analytics skills led to the development of an Online Graduate Certificate Program which is the focus of this article.

16.2 The Innovative Design Approach

16.2.1 *Graduate Certificate in Analytics*

This Graduate Certificate in Business Analytics is a 12 h, on-line credential program designed to give business and non-business graduate students a foundation in (1) the gathering and management of transactional and informational data, (2) statistical methods used in data/business analytics, and (3) the effective application of data mining tools and techniques. Key program components include -

- *Analytics* – foundational analytical & statistical techniques to gather, analyze, and interpret information – "what does the data tell us?"
- *Data* – store, manage, and present data for decision making – "How do I get the data – big data?"
- *Data Mining* – Move beyond analytics to knowledge discovery and data mining – "Now, let's use the data to build models; putting the data to work"

   Students in this program gain knowledge of and hands-on experience with:

- Fundamentals of analytics, decision support, and estimation models

   - Data analysis and business analytics using fundamental statistics
   - Estimation model development using regression and decision trees
   - Forecasting
   - Survey development and analysis
   - Analytics presentation for decision support

- Data management concepts

   - Relational & dimensional data modeling
   - Normalization
   - Structured query language (SQL)
   - Data warehousing & OLAP
   - Data extraction for analysis
   - Database application development

- Data mining
  - Application of analytics to real-world and "Big Data"
  - Data mining: supervised/directed and unsupervised/undirected
  - Data Mining techniques: Linear Regression, Neural Networks, Decision Trees, K-Nearest Neighbor, Logistic Regression, Clustering (Segmentation), and Association Analysis.
  - Web and Text Mining

Students in this program using tools and data sets from SAS (Enterprise Guide, Enterprise Miner), IBM (SPSS Modeler), Microsoft (SQL Server, Analysis Services, Visual Studio, Visio, Excel), Teradata (SQL Assistant), and large scale databases (Sam's Club, Dillard's, Acxiom, Tyson Foods, and other data sets)

16.3   Quality Matters

Quality Matters (2011) provided the umbrella for the overall design of the program. As stated on Quality Matters website, "Quality Matters (QM) is a faculty-centered, peer review process that is designed to certify the quality of online and blended courses. QM is a leader in quality assurance for online education and has received national recognition for its peer-based approach and continuous improvement in online education and student learning." QM identifies eight standards consisting of 63 items (see Table 16.1). This design effort focused on the top 21 items selected from the 8 standards. Within this umbrella guiding the design, other sources were used ensure specific needs of analytics content via online were top quality. Selected practices from Means et al. (2010) evaluation of evidence-based practices in online

Table 16.1 Quality matters standards

Standard:
General Standard 1: The overall design of the course is made clear to the student at the beginning of the course
General Standard 2: Learning objectives are measurable and clearly stated
General Standard 3: Assessment strategies are designed to evaluate student progress by reference to stated learning objectives; to measure the effectiveness of student learning; and to be integral to the learning process
General Standard 4: Instructional materials are sufficiently comprehensive to achieve stated course objectives and learning outcomes
General Standard 5: Forms of interaction incorporated in the course motivate students and promote learning
General Standard 6: Course navigation and technology support student engagement and ensure access to course components
General Standard 7: The course facilitates student access to instructional support services essential to student success
General Standard 8: The course demonstrates a commitment to accessibility for all students

learning were incorporated. The discussion by Anderson et al. (2005) learner paced education led to the adoption of a cohort approach. Salmon's (2002) five stage E-tivities model provided guidance for E-moderating, technical support and amount of interactivity over the steps from (1) access and motivation, (2) online socialization, (3) information exchange, (4) knowledge construction, and (5) development. Finally, Bloom's (1956) taxonomy and Tomlinson and McTighe (2006) played a significant role in preparation of the goals and objectives.

The development process included a partnership between the faculty members and instructional designers. In addition, faculty members for each course worked closely with each other with the guidance of a course designer. Common layouts, learning and assessment approaches are consistent across all four courses. Across all courses are common navigation themes such as (Refer to Appendix I for some examples)–

– Content Areas
– Modular development for enhanced student learning
– Standardized week to week consistency to enhance student accessibility
– A "To Do" list that directly follows the flow of a specific unit or week
– Measurable Learning Objectives for each unit or week
– Specialized Videos and Handouts available for the unit or week
– Assignments, Quizzes, and Exams designed to accomplish and measure learning objectives

16.4 The Four Online Courses

16.4.1 IT Toolkit

Course Description: The IT Toolkit course provides fundamental knowledge and skills in several major areas of information systems in a modular format. For the online Business Analytics Certificate course, we focus on two modules that emphasize the management and use of data in modern organizations – Intermediate & Advanced Spreadsheet Topics, and Relational Databases & SQL.

Goals of the Course: After completion of this course, the student will be able to:

16.4.1.1 Intermediate & Advanced Spreadsheet Topics

• Use modern spreadsheet software to: Describe common benefits and limitations of using spreadsheet software: Explain the basic principles, current applications, and potential uses of "Big Data"

– Import data from multiple sources (files, databases, etc.)
– Manipulate and transform data to conform to specified requirements
– Create simple macros for procedural data manipulation
– Construct meaningful reports and charts from data using multi-dimensional "pivot" features

- Explain basic concepts in data integration
- Demonstrate how visualization can support decision making
- Describe fundamental programming concepts and object-oriented principles

16.4.1.2   Database & SQL

- Explain fundamental concepts related to relational database management systems, data integrity, and data quality

  - Given a set of requirements, determine the types of entities, attributes, and relationships needed to represent the underlying data
  - Construct a conceptual data model using an Entity Relationship Diagram (ERD)
  - Translate a conceptual data model or a user view into a well-structured logical database design using normalization.

- Utilize the Structured Query Language (SQL) to perform operations on data within a database to retrieve desired data using single- and multi-table (join) queries.

16.4.2   Decision Support & Analytics

Course Description: Analytics – information support for the manager-user. An introductory study of systems providing analytics-based information derived from databases within and/or external to the organization and used to support management in the decision making. Application of tools in business analytics, problem solving, and decision making.

Goals of the course – After completion of this course, students will be able to:

- Interpret and present analytics results to management for decision support

  - Understand decision support systems and the role of the different types of systems in today's companies, along with related implementation and design issues.
  - Explain the use of analytics in decision making (scorecards and dashboards)
  - Explain, interpret, and apply the decision analytics results for decision making
  - Communicate analytics results to management

- Utilize commonly used tools of decision analytics
- Apply data analytics and summary statistics concepts to data
- Collect and summarize data for analysis
- Use analytics to make business decisions
- Using analytics, model, and forecast data for business decisions
- Develop and prepare information for decision support scorecards and dashboards using fundamental analytics and business

16.4.3   Data Management Systems

Course Description: This course in database systems is designed to present database concepts as a component of management information systems. The importance of database systems and IT systems in such areas as government, industry, and education, and the growth of theory and applications of these systems to the management decision process underscore the desirability for this course.

The course presents database processing as an essential part of IT systems. General concepts and methods of information systems are be presented with examples in marketing, finance, personnel, and production systems. Since the essence of most information systems is the database system, the database system is the main area of concentration for this course. Physical representation, modeling, commercial systems, and implementations of database systems are presented. Emphasis is given to the database system as the central repository and data warehouse of management information/decision support systems.

Goals of the course: After completion of the course, students will be able to:

- Explain the fundamentals of database management systems,
- Explain the general architecture of database modeling,
- Discuss and comment on the components of a database, as well as the advantages and disadvantages of a database system,
- Explain ethical decision making with its components and be able to further develop an ethical decision making model as an IT professional,
- Explain file organizations and data structures in the context of a database,
- Explain, as well as compare and contrast the database modeling approaches with emphasis on *relational database modeling*,
- Model business problems using the entity relationship modeling approach,
- Answer business questions by obtaining data from a database using SQL and OLAP on MS SQL Server, IBM DB2, and/or Teradata platforms,
- Develop a business system using the relational modeling approach on MS SQL Server, IBM DB2, and/or Teradata platforms,
- Present and further develop specialized contemporary and emerging topics of database processing,
- Explain current data warehousing concepts and uses in business to support decision making,
- Develop and utilize a data warehouse, and
- Explain data mining concepts.

16.4.4   Business Intelligence

Course Description: The purpose of this course is for students to understand the analysis environment encompassed by business intelligence/knowledge management and the related concepts and technologies (such as data warehousing, data mining, etc.) required for successful implementations of such projects. Because of

a lack of generally accepted clarity of the meaning of terms, misunderstanding can occur when different people use a different base reference for these terms. For example, sometimes business intelligence and data warehousing are considered to be the same. Of course, there are gray areas but in this course, the terms are framed such that students should be able to codify and explain, with examples, the terms and concepts in these related topics included within business intelligence and knowledge management. Emphasis is on data mining.

Further, students have the privilege of working with a wide variety of software tools and data in computing environments that are not generally available to most students. Individual and team projects provide opportunities for students to demonstrate understanding and creativity of the concepts covered in class. The course focuses on the concepts and skills needed to develop decision-centric business intelligence and knowledge management applications. This umbrella encompasses a number of other concepts and technologies which are be included in the course.

Goals of the course: *After completing this course, students will be able to:*

- Explain, with examples, the major components of business intelligence
- Present what data mining is and is not
- Demonstrate ability to formulate business problems and transform them into data mining tasks
- Categorize the data mining tasks and methodologies
- Explain conceptually how various data mining techniques work
- Supervised/Directed (Predictive) data mining tasks

  – Estimation
  – Classification

- Unsupervised/undirected data mining tasks

  – Association Analysis
  – Cluster/Segmentation

- Apply appropriate data mining techniques for data mining tasks designed to solve business problems/opportunities
- Explain "Big Data" (including social media) concepts and their value to decision making
- Utilize data mining software to accurately and effectively develop data mining models
- Explain the data mining model output and how it relates to the business problem/ opportunity

16.5 Conclusion

Documentation from many sources confirms increasing demand for analytics for the foreseeable future. Academics are taking varying approaches for preparing students with knowledge and skills for the analytics world. Described herein is an approach

taken by the Sam M. Walton College of Business at the University of Arkansas based on an innovative approach with content incorporating social media analysis. The program started Fall 2013.

16.6   Appendix I: Course Extracts and Samples

Standardized Content Areas:
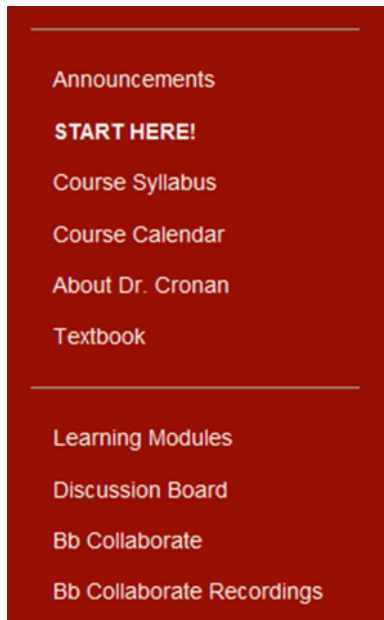
Common "To Do" List:

Measurable Learning Objectives:

Objectives and Introduction:

Learning Objectives - after completing this lesson, you will be able to:

- explain what this introductory analytics course is about and how to complete the course satisfactorily.
- identify and explain Volume, Velocity, and Variety
- explain and give examples and uses of Big Data
- explain business analytics (why, what, how)
- explain management dashboards and scorecards used to support decision making

Week 1 - An Introduction to the class ... an introductory video about the class

Welcome to Decision Support Analytics. This week, we will learn about this course and begin our exploration of the power of analytics. Please watch this short introduction video about the class to get started:

Name: **Course Overview - Bb Navigation**
Duration: 00:09:02

Watch Video

Assignments designed to meet and measure Learning Objectives:

Exercises for this Week:

After reading the Chapters and watching the videos,

- You may work any of the *even numbered exercises* at the end of Chapters 1, 2, and 3; however, be sure to focus on the following exercises --
  - Chapter 2 - 44, 46, 48, 58
  - Chapter 3 - 60, 64, 66, 68, 72
  - the answers are available from the e-text
  - these are optional and for practice
- Datasets for the exercises are available in **Additional Content Resources** (below)

References

American Management Association Study. (2013). *Demand for analytical skills to grow sharply over next five years*. http://www.amanet.org/news/8598.aspx

Anderson, T., Annand, D., & Wark, N. (2005). The search for learning community in learner paced distance education: Or, 'Having your cake and eating it, too!'. *Australian Journal of Educational Technology, 21*(2), 222–241.

Bloom, B. S. (1956). Taxonomy of educational objectives. Published by Allyn and Bacon, Boston, MA. Copyright (c) 1984 by Pearson Education. http://www.elo.iastate.edu/files/2014/03/Quality_Matters_Rubric.pdf.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. New York: McKinsey & Company. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. Technical report*. Washington, DC: U.S. Department of Education.

Quality Matters (QM). (2011). Is a faculty-centered, peer review process that is designed to certify the quality of online and blended courses. QM is a leader in quality assurance for online education and has received national recognition for its peer-based approach and continuous improvement in online education and student learning. http://www.elo.iastate.edu/files/2014/03/Quality_Matters_Rubric.pdf.

Salmon, G. (2002). *E-tivities: The key to active only learning*. Sterling: Stylus Publishing. ISBN 0 7494 3686 7.

Tomlinson, C. A., & McTighe, J. (2006). *Integrating differentiated instruction and understanding by design: Connecting content and kids*. Alexandria: Association for Supervision and Curriculum Development.

Chapter 17
Business Intelligence at Bharti Airtel Ltd

Prabin Kumar Panigrahi

Abstract Bharti Airtel is an early adaptor of business intelligence solution in Indian telecom industry. Over a period of time, the company undertook many IT initiatives. In order to align IT with organization's business strategy, the company has selected and adopted appropriate IT infrastructure and enterprise information systems such as Enterprise Resource Planning (ERP), and Customer Relationship Management (CRM). Subsequently the company implemented Data Warehousing (DW) and Business Intelligence (BI) systems to leverage the IT systems implemented. The company achieved many benefits out of these systems. This Case describes how the company adopted several IT initiatives, leveraged the systems and derived business value by using BI systems.

Keywords Bharti Airtel • Business intelligence • Data warehousing • Customer relationship management

17.1 Introduction

While overseas telecom companies have outsourced tasks to Indian companies, Bharti Airtel is the only Indian telecom company which has outsourced most of its business tasks. A unique feature in Bharti's business strategy was emphasizing customer and product usage data and analytics as a key capability differentiator and retaining that capability in-house and outsourcing all other services to IBM, Erickson and other providers. Using this strategy, Bharti Airtel has spread its operations all over India and has become a dominant telecom company. The company has also drawn international attention. Due to regulatory requirements, an exponential growth

---

This case is developed based on secondary sources and solely as the basis for class discussions. Case is not intended to serve as endorsements, sources of primary data, or illustrations of effective or ineffective management. In some situations opinion of industry experts were also obtained.

P.K. Panigrahi (✉)
Department of Information Systems, Indian Institute of Management Indore,
Prabandh Shikhar, Rau – Pithampur Road, Indore, Madhya Pradesh 453556, India
e-mail: prabin.panigrahi@gmail.com

in its customer base, growing demand for better customer service and competitive pricing, Bharti Airtel implemented several IT systems such as ERP and CRM. Subsequently the company has adopted data warehousing and business intelligence solutions in order to leverage the IT systems that it had adopted. The strategic use of data and analytics helped managers make informed decisions.

17.2 Bharti Airtel Limited

Bharti Tele-Ventures, a family owned telecom business founded by Sunil Mittal, was incorporated as a company on 7th July 1995 for providing telecommunication services. For the first time, Government of India allowed a private company through a bidding process. The company launched its cellular services 'Airtel' in Delhi in 1995 (Bharti Airtel Limited 2012). The company changed its name to Bharti Airtel Ltd on 24th April 2006. Subsequently, it spread its operations in all the telecom circles and with introduction of Unified Access Services (UAS), the company became an all India leading telecom service provider. The number of subscribers crossed 200 million in India and over 260 million in the world by first quarter of 2013 (Bharti Airtel world's fourth largest operator: Report 2012). The company offers mobile, fixed line, broadband, IPTV, Airtel digital TV, Direct-to-Home and enterprise services under the brand 'Airtel' directly or through subsidiary companies. Bharti Airtel provides services in mobile, telemedia and enterprise sectors in different Strategic Business Units (SBUs). The organization structure is given in Fig. 17.1. The company was pioneer in bringing hello tunes, blackberry, iPhone and subsequently M-commerce to India (Bharti Airtel foresees super success on MNP platform 2011). Bharti network covers nearly 87 % of the country's population. The company which was operating in a single circle in 1996 with customer base less than 25 k, USD 17 mn revenue, USD 2.5 mn EBITDA (Earnings Before Interest, Taxes, Depreciations and Amortization), USD 1.4 mn cash profit, and with USD 16 mn market capitalization became the largest integrated private telecom operator in India in 2012 with 252 mn customers, USD 15 bn revenue, USD 5 bn EBITDA, USD 4 bn cash profit and USD 25 bn market capitalization (Bharti Airtel Ltd. India Sustainability Report 2012). In the last couple of quarters of 2012–2013 Bharti Airtel incurred a large debt in market and had negative growth in profits in spite of increase in revenue and number of customers. This trend was due to its acquisition, license fee, spectrum issue and competition. Due to customers carrying multiple connections, rolling of 3G and 4G technology, and expansion in rural sector, the revenue of the company did not increase even if new subscribers were added every month (Ribeiro 2012).

17.2.1 IT Infrastructure at Bharti Airtel

Bharti's telecom business was enabled and supported by various hardware and software in three categories of IT infrastructure i.e. network IT, functional IT and enterprise IT (Bharti Airtel Limited 2012). As soon as the company started it's operations
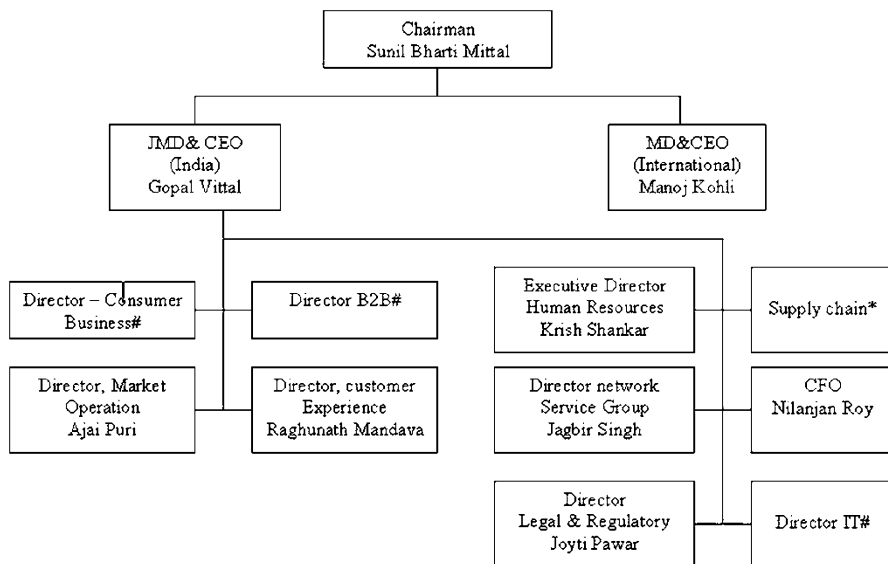
Fig. 17.1 Bharti Airtel Limited: organization structure (Source: Bharti Airtel Website. Accessed 26 June 2013

in 1995, it initiated planning and investment in IT systems (Sharma and Subramanian 2008) in Delhi circle. The company expanded its business in terms of products and services, and also through mergers acquisitions (Neel Mani 2006; The IT Perspective on Bharti Airtel's Incredible Expansion 2012). At the same time customers for different services also grew. Over the period of time Bharti Airtel spread its operations in different circles of the country. The company implemented several functional and enterprise IT systems and required processes as on when required by different units. Each new circle started operating with its own independent WAN (Wide Area Network) with LAN (Local Area Network); network operating center and security layer. In this way, a number of independent disparate business processes, software applications and storage systems were made functional in different circles across the company during 1995–2002. Many of the legacy systems were running with high level of human intervention and not automated. In 2002, the company decided to get rid of its legacy systems, to standardize all customer-facing applications and to integrate various disparate processes so that different categories of users can access the centralized system. As part of the initiative of SBU integration, the entire IT infrastructure migrated to Multi Protocol Label Switching (MPLS), one-to-one communication protocol based services WAN where data, voice and video layers were connected (Neel Mani 2006). The new structure supported all standard applications thereby reducing the costs dramatically. The intra SBU migration process was completed in 2010. The internal information systems for all the three divisions

were put in one platform. The integration resulted a single standard platform for all the functional areas (Neel Mani 2006).

The challenge for the company was to make investment on these resources. On the other hand, investment in IT infrastructure was capital expenditure and would affect the revenue. At the same time, the company was experiencing decreasing Average Revenue per User (ARPU) for mobile services in India although it's ARPU is always remains higher to industry average, and it's Subscriber Base (SB) and Market Share (MS) are higher than other players. ARPU, SB, MS and number of operating circles in India over the period 1995–2013 are provided in Table 17.1. Change in government-mandated pricing policy was one of the factors of declining ARPUs (Bharti Airtel Grows At A Stunning Pace By Keeping Its Focus On The Customer 2008). Bharti Airtel quickly realized that in order to compete and to keep pace with future growth, focusing on its core competency was highly essential. The company decided to outsource all its non-core activities, focus on developing new services, and maintains customer relationships. In July 2004, the company tied up with Ericsson, Nokia and Simens to manage its entire telecom network. The outsourcing contract valued $400 million and Bharti paid the charges based on customer traffic and quality of service (Sharma and Subramanian 2008). Siemens, Nortel, Corning and others became its equipment suppliers for broadband,

Table 17.1 Performance of Bharti Airtel

Year	ARPU-India	ARPU-Bharti Airtel	SB-Bharti Airtel	MS-Bharti Airtel	Circles–Bharti Airtel
1995					1
1996					1*
1997					2
1998					2
1999	1,297				2
2000	1,158				2
2001	814				5*
2002	731		1.35	20.63	9
2003	503		3.07	23.62	15*
2004	412		6.5	19.29	22*
2005	352	406	10.97	21	22
2006	316	343	19.58	21.72	23
2007	280	306	37.14	24.56	23
2008	264	261	61.98	23.74	23
2009	205	201	93.92	23.97	23
2010	131	154	127.62	21.84	23
2011	100	201	162.2	19.99	23
2012	140	189	241.1	24.55	23
2013	130	150	220.3	23.65	23

Notations: ARPU-GSM–Average Monthly ARPU (Mobile Services) | SB in Millions for (GSM+CDMA)| MS in % for GSM+CDMA, Number of Circles (GSM and CDMA, CDMA Services started in India 2001)
Source: TRAI and COAI Reports, Government of India

telephone and enterprise carrier services. In order to provide customer experiences and to make the business scalable, Airtel outsourced its call centers' technology requirements and operations to Nortel and a set of companies such as Hinduja, Mphasis, IBM Daksh, TMT, and Teletech respectively. Similarly the company outsourced management of entire IT infrastructure and related operations to IBM in March 2004 on the basis of revenue sharing. The outsourcing deals were variable costs based on capability usage and revenue. It was a comprehensive $750 m 10-year agreement with IBM. IBM took the ownership, and responsibility of transforming Airtel's key business processes related to IT, and management of IT as a business linked financial model. IT department of Bharti Airtel closely monitored and evaluated the execution plans of partners, Service Level Agreements (SLAs) and performance, and ensured service delivery as per benchmarks (Singh 2011). The company concentrated on understanding customers as well as marketing and branding of product (Govindarajan and Ghosh 2010).

17.2.2   Key IT Systems at Bharti Airtel

Bharti Airtel has implemented and adopted several key IT systems such as Call Data Record (CDR) system, Billing System, Order Fulfill Management System, Unified Geographic Information System (UGIS), Customer Relationship Management (CRM), e-CRM, Data Warehousing and Business Intelligence (BI) System. Transaction Processing Systems (TPS) such as billing, order fulfill management; UGIS and CRM etc. are data and information sources for Data Warehousing and BI system.

As part of planning and investment in IT systems, Airtel automated its customer billing system (Sharma and Subramanian 2008) in Delhi circle as soon as it started its operations. Since inception, Bharti adopted a systematic approach for its data integration and management. The company introduced Call Data Record (CDR) when it started its operation in Delhi circle in 1995 (Sharma and Subramanian 2008). In 1996, Bharti Airtel implemented Oracle Financial, an ERP system that served as the billing system. The system takes CDR data as input and generates billing records automatically. The company wished to create one brand Airtel, across all SBUs and divisions of the company. A commercial off-the-shelf billing system "Kenan" was implemented in 2002 integrating two circles. Subsequently, the mobile division adopted Comverse Kenan in the entire enterprise and provided decentralized billing services to customers in the post-paid sector. In 2011, the company shifted its billing on to a common engine known as convergent billing platform.

In earlier systems various processes for Order Fulfillment System (OFS) were non-standardized across different products and services as well as in different circles. Whenever a customer requests for one particular service, the request travels from a disparate CRM system to a legacy FX billing system. The manual and non-business rule driven conversion of selling to provisioning or billing view led to consume more time, more errors, more tampering and scalability problems. Airtel addressed the issue by implementing order decomposition method (Kumar 2012).

The business rule driven OFS application bridged the gap of e-CRM and FX billing system by integrating both systems and using WebSphere technology solution. The technology enabled standardization, integration and automation of business processes. Customers gained consistent and better experience across touch points. For the company, activation time for customer service as well as cost for servicing a customer reduced (Kumar 2012).

In order to connect most of the villages, cities, railroad and highways of the entire country, Bharti Airtel implemented Unified GIS (Bharti Airtel Unified GIS 2010). The UGIS system stores network and GIS related information. The system was integrated with company's ecosystem of business processes and applications (e.g. CRM), and helped the company in providing networking services, accessibility, and acquiring customers. Tedious manual process of site analysis, feasibility, and maintaining networking inventory was no longer required. Senior management used this as a marketing planning tool in expanding network reach, launching various products and services, and modifying tariff plans. Due to this setup customer service response time was reduced dramatically; tower planning was done more scientifically and complain as well as incident management were improved. As UGIS was integrated with company's internal systems, On-Line Analytical Processing (OLAP) and analytics were applied on location-based information for further analysis. IT partner IBM was the implementation partner whereas Infotech, and Lepton were the system integration partners for the project UGIS (Bharti Airtel Unified GIS 2010).

Bharti started its business with the help of manual systems. At the same time the number of dealers, vendors, and customer base across country started growing significantly. The company was not able to provide centralized customer services and common brand experience anywhere in India. Company found difficulties in resolving customer issues. For example, customers were forced to carry scratch cards for recharging when they move from one location to another. Also they were not able pay their bills anywhere in India. Brand managers became worried about Airtel's brand image (Neel Mani and Shah 2006). Before 2003, for each business process, individual modules such as order processing, sales management; channel management, billing and call center operations had been implemented. There were multiple applications and lacking single view of customer database. It was difficult to know the changing needs of customers as well as consolidated view across various businesses and functions (CRM gives one view of customer facing processes 2009). Airtel decided to implement enterprise level CRM and developed its roadmap for centralized CRM (Customer Relationship Magic at Airtel 2012) for post paid (Neel Mani and Shah 2006). Customer's privacy policy was also another reason of implementing CRM. The company planned to integrate all its business processes across multiple functions with CRM after internal restructuring and process reengineering. As part of this initiative, a WAN with storage area network (SAN) was installed connecting all the major locations as well as a data center at Gurgaon. With the help of technological partners, Oracle CRM Discoverer went on live in 2004 with modules like marketing, planning, campaign management, lead, and sales management etc (Bansal 2010). Subsequently the system was implemented at key customer touch

points of various CRM service operations in all the segments across Pan-India (Dutta 2009). A single point service was provided to all the post-paid customers across India with the help of a centralized CRM. Just after the registration of a new customer, information was made available in CRM system and hence accessible to all front-end systems (Neel Mani and Shah 2006). CRM empowered the managers to know customers' requirements and usage patterns.

Subsequently, taking IBM as technology partner, Airtel implemented e-CRM (Bansal 2010). e-CRM had various features such as online customer support, web interface, customer profiling, and e-billing. The system could able to generate customized bills. A central database of customer related information is created for e-CRM to enable a Pan India service delivery (Airtel offers more features 2002). Customers could able to access their information as part of self-service features and also get answers to most of their queries from the company automatically through call centers. At the same time, the company could able to get feedback from different sources and able to manage the growing number of complains.

With the help of Service-Oriented Architecture (SOA), an IBM Webspere technology (Reaping the rewards of your service-oriented architecture infrastructure 2008), IBM implemented a system that integrated Airtel's customer service applications and related processes. Web, IVR, and SMS – all three major channels were integrated with customer self-service applications. All disparate applications, databases, software services and platforms were integrated using web services available in SOA. Information stored in various legacy systems was integrated with contemporary systems that created an enterprise wide system. Business activity monitoring improved the productivity of managers. IBM SOA provided real time response to customer request (Bharti Airtel manages 110 million subscribers like clockwork with IBM solution 2010).

17.3   Business Intelligence in Telecom Sector in India

In past several years, Indian telecom sector has become more competitive. Due to liberalization and deregulation, the industry got exposed to global business environment besides facing tough competition from domestic competitors. The number of subscribers has grown so as the number of telecom operators. In July 2012, the wireless subscriber base was 913.49, and tele-density was 75.21 where as for wireline, the subscriber base was 31.33 and tele-density was 2.58. The total subscriber base was 944.81 and tele-density is 77.79. This made the market customer centric rather than product centric. At the same time, telecom policies have also been changed over period of time that include concept of virtual network operator, and benchmarks for quality of service. The industry was facing decreasing ARPU and revenue, and increasing debt (Bharti Airtel Company Profile 2009). The challenges any Indian telecom player faces are churn, quality of service, and customer growth. Instead of technology, customer service is the driving force for customer retention. In future the growth drivers for telecom service providers in India would be new

technology, new services like 3G, BWA, and better devices (A brief report on Telecom sector in India 2012). The company must understand customers' needs so that they can take proactive informed decisions and remain competitive. Decision-making requires subject-oriented views rather than transaction view. Data ware-housing and BI helps the company in addressing the above issues and challenges. The comprehensive BI solution sits on core IT systems of the organization and is integrated into company's business processes. The system provides right knowledge to the right people at right time. This allows management across network opera-tions, line of business, call center and corporate to understand customers and their behaviour, plan, monitor and visualize performance.

In this context customer data become most important asset for the organization. The volume of data generated in a telecom company grows exponentially as span of operations increases, new services are added and number of transactions increase. Several petabytes of data are generated on a regular basis. There are two main sources of data for a telecom company: Main Switching System (MSC) and Customer Relationship Management (CRM). Using a mediation software the CDR system (Table 17.2) integrates these two sources of data by pulling in call records (CDRs) from the MSC and customer details from the CRM data base for post paid customers for whom the bill is to be generated on monthly basis. A billing database updated. It then carries out the rating of the CDRs based on the duration and desti-nation of the calls and customer category. Based on these inputs, it generates the amount that is to be billed to the customer for that call. It then aggregates the billing data over the period of billing cycle and this aggregated data is used for generation of customer bills by the CDR system.

17.4 Business Intelligence and Data Warehousing at Bharti Airtel

17.4.1 Data Warehousing

Over the period of time, Bharti Airtel maintained various transaction data in data-bases across business units and lines of business. Transaction processing systems provide the transactional view of operations and processes such as customer invoic-ing, stock control, complain management and bill payments. Although a company's operational efficiency could be improved, it is not possible for the company to get subject-oriented view of business such as customer, product, and sales at top enter-prise level. The company must integrate the data from various transaction systems into one data warehousing so that analytics, OLAP and various reporting tools could be used to get actionable information and insights. A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data. It stores cur-rent as well as historical data and is organized around subjects such as customers, products and sales. It supports reporting and query tools. Database or Online-Real Time Systems (OLTPs) are designed for efficiency where as data warehouse is designed for decision-making.

Table 17.2 Data descriptions of telecom data

Data type	Descriptions	Information about	Characteristics	Source/Generated
Call detail data	Customers' Calls that traverse the telecommunication networks	Originating and terminating phone numbers/address, the date and time of the call, time the call got connected, successful or failure, message, roaming location details and the duration of the call	Real-time, Time Series, Individual Event	Generated at Main Switch Center (MSC) in raw form. Processed and used for billing for CDR rating
Network data	Status of the hardware and software components in the network: availability of network elements	Timestamp, a string that uniquely identifies the hardware or software component generating the message and a code that explains why the message is being generated. Used for maintenance and monitoring performance of the network	Real-time, Time Series, Individual Event	Generated at MSC
Customer data	Registration data when a new customer takes a mobile connection at the time of activation	Name, billing address, mobile number, type of connection-postpaid/ pre-paid, service plan, contract details-email/phone	Not Real time. Supplemented with data from external sources, such as credit reporting agencies	Generated at CRM module at the time of new mobile connection registration and then used in MSC when customer activation
Billing data	State of profile, plan and usage information of customers	Usage data, service plan, customer type, payment history, roaming details, value added services	Not real time, typically made available once per month	Available at billing system for rating of CDR based on CDR pulled from MSC

Source: Weiss (2005)

In 1997, the company reviewed the existing systems and felt that the organization was maintaining different application systems for different business area. The transaction data was available in disparate IT systems for which seamless flow was not possible and hence demand for information was not fulfilled (Sharma and Subramanian 2008). Managers were not able to get a unified view of customer, sales, and other subjects. Duplication in data and process created many problems such as system failure, inconsistency, and partial view of customer information. In the same year, company realized the need for enterprise data integration and

management and planned for implementing data warehousing system. It believed that data warehousing system would help them in integrating their existing data available in disparate systems and pave the way for digging and exploring rich customer information. Data warehousing was considered as a strategic intervention that would help the managers in taking informed decisions.

Bharti started planning and implementing data warehousing early in the year 1999 taking help from Telecom Italia for a short period of time in the requirement phase (Sharma and Subramaniam 2008). During that time different business issues and opportunities were looked into in order to get a holistic view of information needs of the company. Accordingly the DW was designed and customized. The company selected and implemented Oracle Express as Enterprise Data Warehouse (EDW) out of Oracle Express and SAS. Subsequently due to some issues such as scalability (increase in number of circles and customer size), changes in reporting requirement, changes in control systems and processes (e.g. billing cycles) and data integration across systems, Airtel implemented Teradata, a full-scale, and high-end data warehousing system (Sharma and Subramanian 2008). In 2004, Teradata India, tied up with Bharti to provide data warehousing and business intelligence platform (Bharti Airtel manages 110 million subscribers like clockwork with IBM solution 2010; Sharma and Subramanian 2008). Major data sources of data warehouse are network systems, CRM/e-CRM, Billing, Service related, GIS, and CDR systems.

After implementation, the internal team monitored the system closely and uncovered various issues. One of the positive aspects of implementation was that, the employees accepted and appreciated DW intervention. Due to DW system, various anomalies, and deviations in various business processes were discovered. Accordingly key business processes were identified, streamlined and re-engineered. Also new innovative processes were added into the system. The corresponding policies and workflows were scrutinized and rectified. The reports generated from data warehouse triggered changes in internal control systems and key business processes for such as billing process (Sharma and Subramanian 2008).

17.4.2   Business Intelligence

Business Intelligence (BI) is a strategy by which organizations use data assets to understand customers better and make informed and proactive decisions. It is a set of software and solutions for gathering, consolidating, analyzing mostly internal data and providing access to information in a way that helps enterprise users to monitor business activity and to take informed business decisions. BI is a broad term that includes query, reporting, Online Analytical Processing (OLAP), Data Warehousing (DW), data mining, forecasting and statistical analysis. The solution is used to leverage the data that are available in organization's enterprise systems such as billing, and CRM. Overall decision-making and efficiency (IT and business side) are positively impacted because of business reporting and better data analysis. Business intelligence helped Airtel in collecting, systematizing, analyzing and

communicating information used in business (Press enter for business intelligence 2002). In India, where the ARPU is the lowest in the world, telecom companies earn profit by cross-selling products and services and by retaining existing customers.

> Data analysis is essential for telcos to run the business on a daily basis as it helps create a differentiation, says Bharti's Gangotra (Mahalingam 2012).

Airtel considered BI as a strategy that was enabled by data warehousing. The system worked as cross-functional and presented information in different forms as required by various categories of department users at different levels of the hierarchy. The BI system empowered Bharti Airtel executives in taking informed decisions about different aspects of business. The company implemented IBM Cognos as BI system in 2004 as part of the outsourcing deal. Cognos assisted Airtel managers in getting insights using data warehouse in the form of reports, scorecards, and forecasts (Power your business with intelligence, IBM 2012).

When asked about the BI mission of any company, Rammurthy, BI consultant said,

> BI provides a single platform with sound analytical support. The system provides actionable information to managers that help them in managing performance through planning, and decision-making, and monitoring business activity using tools such as balance scorecard, report, dashboards, and alerts. The mission of BI of a company is to provide an integrated view of the business, market and customer thus enabling the company to offer customized and relevant offerings to its customers. Using predictive modeling and data mining, the company is empowered to take 'informed' business decisions which in turn bring value to business

The Business Intelligence system at Bharti Airtel integrated information across all the important information sources. The system acted as a single platform supporting various analytics required for decision support. BI architecture of a telecom company is depicted in Fig. 17.2.

17.5   Leveraging IT Capability by Using Business Intelligence

Implementation and adoption of appropriate IT infrastructure in terms of hardware and software paved the way for leveraging IT capability at Bharti Airtel. Business units provided the complements. Bharti Airtel started adopting IT culture as early as 1999. Using the SAS, an analytical tool during 1999, the company discovered that a number of subscribers in pre-paid sector in Delhi were not locals but visitors who subscribed to Airtel back in their hometowns. Based on this active information, company launched a regional roaming service for pre-paid subscribers. Similarly as a churn management tool, the company used SAS in Delhi circle and slashed down the churn rate to 2 % from 3 % (Press enter for business intelligence 2002). The managers studied the call drop pattern of customers and discovered that the wrong tariff plan was one the key reasons of churn.
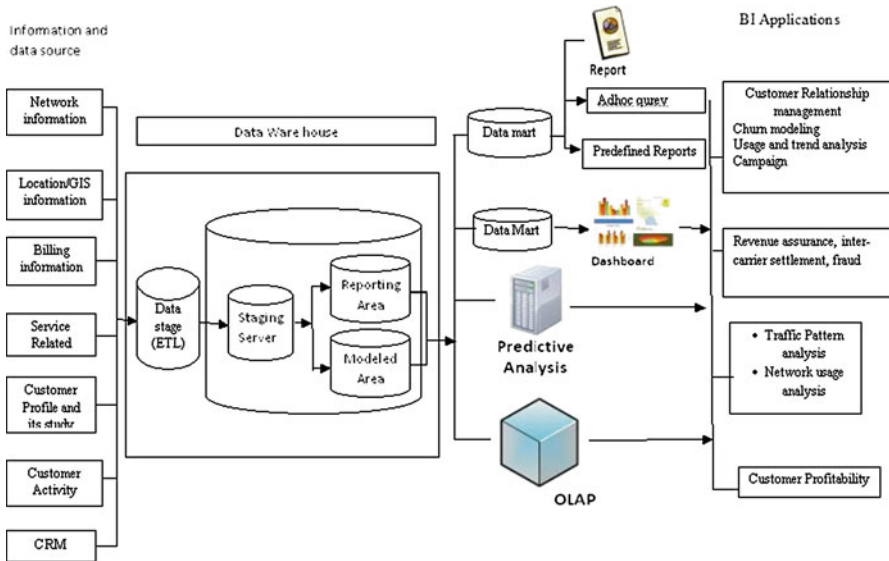
Fig. 17.2 Business intelligence (Source: Adopted from Transforming Telecommunications Business Intelligence 2004; Vodapalli 2009)

An example of how the solution helps is with respect to tariff plans. We were able to re-align tariff plans best suited to specific customers and rationalise their billing cycles. Says Mohit Deora, chief marketing officer, Airtel (Delhi) (Press enter for business intelligence 2002).

Integration of SOA services with BI application enhanced the Business intelligence capabilities of the company (Daftari 2007). Data collected through IBM SOA reusable services were integrated with SOA dashboards and IBM Cognos, the business intelligence application. The integration empowered Airtel's employees to perform various types of analytical processing using OLAP thereby increasing the productivity. OLAP is a BI tool. It provides a 'front-end" to view data in data warehouse. It has powerful querying and data analysis environment such as slicing, dicing, roll-up, drill-down, and pivoting. With the help of fast querying and analysis power, it supports decision-making and analytical activities. Due to BI, the analytical power of CRM improved significantly. Adopting a soft-scaling approach, the managers applied analytics on data empathetically considering risks and opportunities for taking decisions (Agarwal and Weill 2012). While understanding and interpreting various performance measures used for monitoring day-to-day activities, senior executives used contextual and evidence-based data such as structured conversations, word of mouth, emails, text messages, and reports (Agarwal and Weill 2012).

With the help of BI system, Airtel managers (Bharti Airtel manages 110 million subscribers like clockwork with IBM solution 2010) monitored different KPIs

using dashboards and alerts. Dashboards display actionable information related to any specific KPI in a simple way so that managers could take proactive decisions. With this a manager could quickly understand any concern or deviation, trend, threshold crossed, unachievable target and other relevant indicators. Any deviation from assigned KPI value is signaled through automatic alerts on dashboards such as emails, graphics, or animations that draw manger's attention. When business intelligence combined with alerts on the dashboards, the performance of managers increased significantly. A senior manager of BI project of a telecom company said,

> Considering organization's strategy and balanced scorecard framework, the board and executive committee of a company decides a set of financial and non-financial metrics, known as Key Performance Indicator (KPI) by which the company can monitor and evaluate strategic performance of organization towards its vision and long-term goals. KPIs help in assessing the present state of business and prescribes course of actions. Business activity monitoring can be done by monitoring KPIs on a real-time basis.

The benefits of BI systems to Bharti Airtel were many (Sharma and Subramanian 2008). Different users could able to get reports automatically. The information provided insights about various discrepancies or errors in reporting for e.g. between daily usage report from crude telecom switch data and minutes of usage from billing data that was provided by data warehouse. This helped the manager in taking data-oriented decision-making, and also in changing internal control systems. The BI system provided one version of truth as all the departments source data from one data warehouse at same time. The managers could able to segment customers and market the products to the targeted customers (Daftari 2007). The system users could able to get insights of customer behaviour patterns. The managers started using models that helped them in prediction. The financial performance of the company is summarized in Table 17.3.

The value of business intelligence was significant. The source and destination of the calls with details provided intelligence to various departments. Similarly the phone calls from other operators that are routed through the company network was very useful information. Airtel used BI system in identifying high-value customer segments and acquire customers through a product-segment fit (Daftari 2007). Similarly the company used BI and associated tools for running various campaigns (Mahalingam 2012). The information regarding calling pattern of each customer, and type of usage helped in initiating programs in some segment, among existing customers. Business intelligence was also integrated with centralized CRM of the company (Neel Mani and Shah 2006).

In 2007, using the available data, the marketing and communication department found two important segments i.e. 5 % of total subscribers contribute 10 times of company ARPU referred as high end users or achievers and one third of user base who were in the age group 18–35 spend much more than others on VAS (e.g. no of messages) referred as high adopters of VAS or "funsters". Instead of segmentation based on demographic profile, the department segmented its customers based on behaviour and purchase pattern. In order to leverage this knowledge and to get more revenue, company rolled out different set of services, kept different priority cus-

Table 17.3 Bharti Airtel financial performance

Year	Q	R	At	B	C	F	G	D	A(I)	L	M	N	O	P
1996						0	NA	NA				0		
1997	4.79	0.33				48.4	NA	0.09	0	0.07		2.74	5.27	15.97
1998	1.74		0.17	0.14		54.11	NA	0.05		0	−1	0.77	1.11	
1999	3.58		0.53	0.23		.30	NA	0.02		0		1.78	2.67	
2000	8.79		1.65			5.14	NA	0.03		0		4.49	6.36	
2001	26.56		6.42			0.67	NA	0.05		0		9.69	24.32	
2002	56.95		8.03	1.99		0.2	593.08	0.02		0		20.28	53.72	
2003	41.38	16.69	1.78	2.73		0.3	995.26	0.01	0.27	0.4		21.53	37.78	2.26
2004	41.53	29.17	1.71	2.91		0.58	1,302.83	0.01	0.16	0.7	0.75	23.86	37.66	1.29
2005	6,101.6	8,094.9	152.6	771.55	8.4	14.57	31.67	0.83	0.12	1.33	276.5	4,625.5	4,970.9	0.61
2006	8,390.9	11,231	377.64	997.47	8.4	17.76	38.86	0.66	0.12	1.34	0.39	6,423.5	6,850.5	0.61
2007	126.95	17,852	632.84	1,282.2	8.4	22.4	35.88	0.75	0.11	1.41	0.59	9,707.5	102.33	0.57
2008	18,131	25,704	929.28	1,651.7	8.4	23.96	25.11	0.74	0.1	1.42	0.44	13,824	14,877	0.58
2009	23,948	3,4014	1,169	2,062.7	8.4	22.42	15.34	0.73	0.1	1.42	0.32	18,882	20,650	0.61
2010	25,523	35,610	1,053.3	2,495.3	48	25.59	12.57	0.64	0.1	1.4	0.05	20,075	21,536	0.6
2011	28,978	38,018	1,042.9	2,404.6	611	20.04	17.59	0.55	0.11	1.31	0.07	22,936	24,395	0.64
2012	33,350	41,604	1,258.7	2,565.5	837	13.57	22.32	0.5	0.11	1.25	0.09	26,683	27,467	0.66

Notation: Q: Cost of sales (Cr); R- Sales (Cr); At: IT/ITES & other professional services (Cr); B: Computers and it systems, gross (Cr); C: Software, gross (Cr); F-PAT as % of total income; G- PE on BSE (Ratio); D: Total income/Total assets net of misc exp and reval (times); L- sale/cost of sales; M- %sales growth; N – cogs; O- operating expenses; P- oe/sales

Source: CMIE Database for the period 1996–2012. The table is prepared by the author

tomer relation mangers, collaborated with HTC and Blackberry (for high-end handsets), redesigned its WAP web portal, and offered mobile search options (tied up with Google) (Daftari 2007).

Using BI and analytics, the company analyzed usage and recharging patterns of its customers in order to understand its customers better, to cross-sell its products and to create differentiation. BI and analytics helped the managers in creating thousands of campaigns on a monthly basis for e.g. campaign '"My Airtel My Offer". Similarly on a daily basis, the managers analyzed the data and created customized plan that increased the usage (Mahalingam 2012).

> "We analyze usage and recharging patterns with the help of cutting edge technology," says Amrita Gangotra, director-IT, India and South Asia, Bharti Airtel. "This helps us understand our customers better and allows us to cross-sell our products more effectively. We have been able to increase usage based on the offers we've made", says Gangotra (Mahalingam 2012)

Using the available data, a new product 'Airtel Money' was experimented in 2011 for emerging markets. The management and board of the company found certain trends in the business such as increase in use of smart phones, saturation of mobile penetration in urban areas. They leveraged this information and launched open wallet like service called "Airtel Money" with the help of technical partner Infosys (Bharti-Airtel Abridge Annual Report 2012).

The company provided an opportunity to its customers for basic banking services using mobile phones through existing telecommunication network. As the company expected several changes in requirement and its corresponding development, instead of adding to and modifying existing IT systems, it started altogether a new IT system to implement this product and it's expected changes in future. The company thought of integrating both the IT systems in future once this system gets established. Finally, the company launched the product through 'Airtel M Commerce" (Roy 2011). Besides understanding individual usage pattern of users, Airtel has started using predictive model (Shinde 2011).

With reference to various parameters of Quality of Service (QoS), Bharti Airtel met the required benchmarks set by regulatory authorities both for wireless and wire line services. The customer satisfaction index of the company continuously improved and remained satisfactory. The index for Q4 2010-11 to Q4 11-12 was in the range of 85–88 for post-paid and 90–92 for pre-paid (Bharti Airtel Ltd. India Sustainability Report 2012).

17.6   What Next?

High churn ratio (8–9 %) is one of the major issues in Indian telecom industry (Bharti Airtel foresees super success on MNP platform 2011). Although the churn rate in case of Airtel is much less than that of industry, the recent Mobile Number Portability (MNP) has further increased (A brief report on Telecom sector in India 2012) the chance of increase in churn rate in the company. Before MNP, churn modeling was difficult and inaccurate as the estimation was based on addition and deletion of

subscribers in two consecutive years. Due to MNP, it would be easier to find churn rate by using BI tool (A brief report on Telecom sector in India 2012). MNP provided the customer wider choice and switch option for better service. At the same time recent Telecom Regulatory Authority of India (TRAI) norms on activation made customer churn difficult (Bhupta 2013).

A new challenge Telecom service provider would face is big data and its management. In order to discover customer insights more accurately, not only the structured data that are available inside the company databases is required, but also the unstructured data such as video, voice, emails, and online reviews available on social media, phone devices and other sources are essential. In 2010 Airtel started its customer care service "Airtel Presence" via Twitter, Facebook, and Email for its mobile, landline, broadband or DTH. The company intended to take care of the new generation customers in a better way. During the same time, In 2010, the company launched 'Airtel Blog', a voice-oriented blog service in the domain of social networking for its mobile customers. Within three months, it attracted more than two millions users (Airtel Presence – Customer Care Service via Twitter, Facebook, Email 2010).

> "Perhaps one of our fastest growing Value Added Services in our stable in recent times, Airtel Blog is taking the concept of micro-blogging to an entirely new level," said Atul Bindal, President-Mobile Services, Bharti Airtel. "As a nation, we love to verbalise and reach out to our social network of friends, families or pursued interests and Airtel Blog enables this added intimacy on the customer's mobile phone. Some have called Airtel Blog 'Twitter with a voice' but we are simply calling it a game changer (Airtel Presence – Customer Care Service via Twitter, Facebook, Email 2010)."

Similarly the company has implemented IVR services that allow customers to interact with service systems. Automatic analysis and mining of feedbacks available in non-English Indian languages would a big challenge for the company. In this big data era, companies must collect, and analyze the relevant information to get insights (Mahalingam 2012). Currently telecom companies in India analyze their data by taking a sample. In this process, some information is lost. Also it consumes a lot of time in processing. Traditional BI fails to do complex analysis of data that are in terabyte and petabyte at CDR level on a real-time basis. The performance of the system deteriorates. The organization needs a system (e.g. Netezza of IBM), which is capable of handling big data without sampling and summarizing. Certain analysis cannot be done if history is lost or data is summarized and then analyzed e.g. the cause and effects of network events, holiday calling patterns (Transforming Telecommunications Business Intelligence 2004).

Airtel has been using BI and getting standard and ad-hoc queries. In future the company expects to apply real-time BI and prediction such as churn. BI would be used to identify VAS customers (all operational services beyond the traditional voice calls) to drive ARPU and target marketing (Sharma and Subramanian 2008). On future direction of BI, industry consultant Rammurthy remarked:

> In India we need a real-time operational BI that should be used both as frontline tool and analytical tool in business applications. The days are past where focus of DW and BI system was on querying and reporting based on past i.e. what happened. The focus must be extended to predictive analysis and advanced analytical tools to enhance proactive decision-making.

Currently in Indian companies, simple tools and techniques of BI such as query tools, and OLAP are being used for realizing business value of data warehouse and BI systems. The users are not able to get the insight, rules, or intelligence directly. Depending on the understanding level, experience and cognitive level, users discover insight from the data by using these preliminary tools. There is no or limited use and application of data mining tools and predictive analytics. Data mining techniques use complex mathematical formulae to get various patterns from the database. Due to the huge volume of data, memory and processing power limitations, application of advanced analytical tools are very much limited. Related queries take considerably significant time to provide results.

Lack of fully integration of customer facing applications and manual intervention is another challenge for Indian telecom industry. In India, there is also a practice of manipulating call drop and customer service data to avoid actions and simultaneously to be rewarded. This led to poor quality of business intelligence output. When asked to a BI consultant in telecom sector regarding various issues and challenges in a telecom company implementing Business Intelligence, he remarked,

> In case of low performance of business intelligence project, simply blaming software and data is not enough. Unless the BI information is not converted to actions by managers BI cannot be leveraged.

Biography

Prabin Kumar Panigrahi   a Fulbright-Nehru Scholar, is an Associate Professor of Information Systems at Indian Institute of Management Indore, India. He received his PhD from Indian Institute of Technology, Kharagpur, India after completing his Post Graduate in Computer Applications from Regional Engineering College, Rourkela, India. His current research interests include Business Value of Information Systems, E-Governance, and E-Learning. He has international as well as national research publications in refereed journals. He is also a recipient of IIE-CIES (USA) OLFT and Shastri Indo-Canadian Institute STSG awards. He was member of Indo-British Higher Education Link and a certified SAP Global Solution Consultant.

References

A brief report on Telecom sector in India. (2012). Corporate Catalyst India, August. http://www.cci.in. Accessed 2 Nov 2012.

Agarwal, R., & Weill, P. (2012). The benefits of combining data with empathy. *Sloan Review, MIT*, September 18.

Airtel offers more features. (2002). The Hindu Business Line, June 27. http://www.thehindubusinessline.in. Accessed 12 Oct 2011.

Airtel Presence – Customer Care Service via Twitter, Facebook, Email. (2010). November 22, http://teck.in. Accessed 1 Oct 2012.

Bansal, A. (2010). Customer Relationship Management (Airtel). http://www.slideshare.net/ashish1.bansal/customer-relationship-management-airtel. Accessed 26 Nov 2012.

Bharti Airtel Company Profile. (2009). Data Monitor, 21st August. http://www.datamonitor.com. Accessed 12 Nov 2012.

Bharti Airtel foresees super success on MNP platform. (2011). Money Control, Jan 24. http://www.moneycontrol.com/news/business. Accessed 18 Oct 2012.

Bharti Airtel Grows At A Stunning Pace By Keeping Its Focus On The Customer. (2008). January, IBM UK. http://www.ithound.com/abstract. Accessed 14 June 2012.

Bharti Airtel Limited. (2012). http://www.123helpme.com/. Accessed 2 July 2012.

Bharti Airtel Ltd. India Sustainability Report. 2011–12, Airtel, 2012.

Bharti Airtel manages 110 million subscribers like clockwork with IBM solution. (2010). IBM. 31st March, http://www-01.ibm.com/software/success. Accessed 13 Oct 2011.

Bharti Airtel Unified GIS. (2010). PC Quest, June 2. http://pcquest.ciol.com. Accessed 2 Oct 2012.

Bharti Airtel world's fourth largest operator: Report. (2012). *Business Today*. October 21.

Bharti-Airtel Abridget Annual Report. (2012). http://www.airtel.in. Accessed 28 Nov 2013.

Bhupta, M. (2013). Bharti Airtel may return as leader of the pack in FY14. *Business Standard*, January 10.

CRM gives one view of customer facing processes. (2009). Bharti Airtel., NWN News Network., June 01. http://www.informationweek.in/archive/. Accessed 20 Oct 2011.

Customer Relationship Magic at Airtel. (2012). http://www.slideshare.net. Accessed 20 Oct 2012.

Daftari, I. (2007). Airtel to express itself from four platforms. *The Economic Times*, September 13.

Dutta, R. (2009). Bharti Airtel streams implementation with Oracle E-business suite and discover. Oracle. June. http://www.oracle.com/au/products/database. Accessed 20 Aug 2011.

Govindarajan, V., & Ghosh, A. (2010). Reverse innovation success in the telecom sector, May 12. http://blogs.hbr.org/cs. Accessed 27 Nov 2012.

Kumar, V. (2012). Bharti Airtel Improves Order Management, Developer Works, IBM, June 29. https://www.ibm.com/developerworks. Accessed Aug 11 2012.

Mahalingam, K. (2012). Analyse this. *Business Outlook India*, May 26.

Neel Mani, R. (2006). Lesson learns from vertical. *Real CIO World, 2*(01), 42–46.

Neel Mani, R., & Shah, S. (2006). Customer relationship magic. *Real CIO World, 1*(11), 32–41.

Power your business with intelligence, IBM. (2012). http://www.07.ibm.com/sg/smartersystems-now/intelligence.html. Accessed 27 Nov 2012.

Press enter for business intelligence. (2002). *Economic Times, Indian Times*, December 12.

Reaping the rewards of your service-oriented architecture infrastructure. (2008). IBM Global Services. September. ibm.com/SOA/services. Accessed 2 Oct 2012.

Ribeiro, J. (2012). Bharti Airtel revenues increase, profits down by 37%. *Computer World*, August 8.

Roy, D. (2011). How Airtel gave India its first open-wallet mobile service, A case study on Mobility in Telecom. *Computer World*, July 24.

Sharma, N., & Subramanian, S. (2008). Data warehousing as a strategic tool at Bharti Airtel. ISB, Case No. CS-08-001, July 2.

Shinde, S. (2011). Adding intelligence to businesses. *Business-Standard*, Mumbai July 13.

Singh, R. (2011). Successful innovations happen at the intersection of neighboring industries, September 15. http://voicendata.ciol.com/content/speakh. Accessed 2 July 2012.

The IT Perspective on Bharti Airtel's Incredible Expansion. (2012). *A case study on Applications in Data Warehousing*. http://www.cio.in/case-study. Accessed 2 Oct 2012.

Transforming Telecommunications Business Intelligence. (2004. Real-time, comprehensive analyses for proactive business decisions white paper, Netezza Corporation, pp. 1–10.

Vodapalli, N. (2009). Critical success factors of BI implementation, Master's Thesis Report. IT University of Copenhagen. http://www.itu.dk/~navvod/CSFsOfBIimpl.pdf. Accessed 26th November 2012.

Weiss, G.M (2005). Data mining in the telecommunications. In O. Maimon, L. Rokach (Eds.), Data mining and knowledge discovery handbook: a complete guide for practitioners and researchers, Kluwer Academic Publishers, pp. 1189–1201.

Index