Springer Proceedings in Complexity

Lorna Uden
Leon S. L. Wang
Tzung-Pei Hong
Hsin-Chang Yang
I-Hsien Ting *Editors*

The 3rd International Workshop on Intelligent Data Analysis and Management

Springer

Springer Proceedings in Complexity

For further volumes:
http://www.springer.com/series/11637

Lorna Uden · Leon S. L. Wang
Tzung-Pei Hong · Hsin-Chang Yang
I-Hsien Ting
Editors

The 3rd International Workshop on Intelligent Data Analysis and Management

Springer

Editors
Lorna Uden
School of Computing
Staffordshire University
Stafford
UK

Leon S. L. Wang
College of Management
National University of Kaohsiung
Kaohsiung
Taiwan, R.O.C.

Tzung-Pei Hong
Department of Computer Science and Information Engineering
National University of Kaohsiung
Kaohsiung
Taiwan, R.O.C.

Hsin-Chang Yang
National University of Kaohsiung
Kaohsiung
Taiwan, R.O.C.

I-Hsien Ting
Department of Information Management
National University of Kaohsiung
Kaohsiung
Taiwan, R.O.C.

ISSN 2213-8684 ISSN 2213-8692 (electronic)
ISBN 978-94-007-7292-2 ISBN 978-94-007-7293-9 (eBook)
DOI 10.1007/978-94-007-7293-9
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013944207

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Data analysis is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver, and enhance the value of data and information assets. Data analysis and data management both have multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains. Intelligent Data Analysis and Management (IDAM) examines issues related to the research and applications of Artificial Intelligence techniques in data analysis and management across a variety of disciplines. It is an interdisciplinary research field involving academic researchers in information technologies, computer science, public policy, bioinformatics, medical informatics, and social and behavior studies, etc. The techniques studied include (but are not limited to): Data visualization, data pre-processing, data engineering, database mining techniques, tools and applications, evolutionary algorithms, machine learning, neural nets, fuzzy logic, statistical pattern recognition, knowledge filtering, and post-processing, etc.

On June 8, 2012 the IDAM was first held in College of Management, National University of Kaohsiung and the second IDAM was held on May 17, 2013. The third IDAM gathers people from previously disparate communities to provide a stimulating forum for exchange of ideas and results. We invite academic researchers (in information technologies, computer science, business and organizational studies, and social studies), as well as information technology companies, industry consultants, and practitioners in the fields involved.

The IDAM proceedings consist of 15 papers covering different aspects of Intelligent Data Analysis and Management. Authors of the papers come from many different countries such as Australia, India, Korea, Singapore, and Taiwan.

We would like to thank our authors, reviewers, and program committee for their contributions and the National University of Kaohsiung for hosting the conference.

Without their efforts, there would be no conference or proceedings.

Kaohsiung, Taiwan, September 2013

Lorna Uden
Leon S. L. Wang
Tzung-Pei Hong
Hsin-Chang Yang
I-Hsien Ting

Organizations

Conference Chair

Prof. Lorna Uden—Staffordshire University, UK

Program Chairs

Prof. Leon S. L. Wang—National University of Kaohsiung, Taiwan
Prof. Tzung-Pei Hong—National University of Kaohsiung, Taiwan

Local Chairs

Prof. Hsin-Chang Yang—National University of Kaohsiung, Taiwan
Prof. I-Hsien Ting—National University of Kaohsiung, Taiwan

Organization Committee

Prof. Chian-Hsueng Chao—National University of Kaohsiung, Taiwan
Prof. Han-Wei Hsiao—National University of Kaohsiung, Taiwan
Prof. Ying-Feng Kuo—National University of Kaohsiung, Taiwan
Prof. Hsing-Tzu Lin—National University of Kaohsiung, Taiwan
Prof. Yu-Hui Tao—National University of Kaohsiung, Taiwan
Prof. Kai Wang—National University of Kaohsiung, Taiwan
Prof. Chen-Hsing Wu—National University of Kaohsiung, Taiwan
Prof. Shu-Cheng Yang—National University of Kaohsiung, Taiwan
Ms. Ming-Jun Chen—National University of Kaohsiung, Taiwan

Program Committee

Prof. Ajith Abraham—Machine Intelligence Research Labs, USA
Prof. Damminda Alahakoon—Monash University, Australia
Prof. Chien-Chung Chan—University of Akron, USA
Prof. Bo-Rong Chang—National University of Kaohsiung, Taiwan
Prof. Ping-Tsai Chung—Long Island University, USA
Prof. Seng-Cho Chou—National Taiwan University, Taiwan
Prof. Saman Halgamuge—University of Melbourne, Australia
Prof. Liang-Cheng James Huang—Academia Sinica, Taiwan
Prof. Han-Wei Hsiao—National University of Kaohsiung, Taiwan
Prof. Hui-Yin Hsu—New York Institute of Technology, USA
Prof. Yada Katsutoshi—Kansei University, Japan
Prof. Tsau-Young Lin—San Jose State University, USA
Prof. Wen-Yang Lin—National University of Kaohsiung, Taiwan
Prof. Pawan Lingras—Saint Marys University, Canada
Prof. Victor Lu—St. John's University, USA
Prof. Javier Bajo Perez—Polytechnic University of Salamanca, Spain
Prof. James Tan—SIM University, Singapore
Prof. Ted Teng—Stony Brook University, USA
Prof. Shusaku Tsumoto—Shimane University, Japan
Prof. Da-Jin Wang—Montclair State University, USA
Prof. Chen-Shu Wang—National Taipei University of Technology, Taiwan
Prof. Shiang-Kwei Wang—New York Institute of Technology, USA
Prof. Jierui (Jerry) Xie—Oracle, USA
Prof. Hsin-Chang Yang—National University of Kaohsiung, Taiwan
Prof. Qiangfu Zhao—University of Aizu, Japan
Prof. Haibin Zhu—Nipissing University, Canada
Prof. Mihai Horia Zaharia—Gheorghe Asachi Technical University, Romania

Contents

1 **An Information Quality (InfoQ) Framework for Ex-Ante and Ex-Post Evaluation of Empirical Studies** 1
Galit Shmueli and Ron Kenett

2 **Memory-Aware Mining of Indirect Associations Over Data Streams** 15
Wen-Yang Lin, Shun-Fa Yang and Tzung-Pei Hong

3 **Graph-Based Batch Mode Active Learning** 27
Cheong Hee Park

4 **One Pass Outlier Detection for Streaming Categorical Data** 35
Swee Chuan Tan, Si Hao Yip and Ashfaqur Rahman

5 **Measuring of QoE for Cloud Applications** 43
Yu-Hui Tao, Yu-Lung Wu, Chi-Jui Chang and Chi-Wen Chang

6 **Mining Weighted Partial Periodic Patterns** 47
Kung-Jiuan Yang, Tzung-Pei Hong, Yuh-Min Chen and Guo-Cheng Lan

7 **Edge Selection for Degree Anonymization on K Shortest Paths** 55
Shyue-Liang Wang, Ching-Chuan Shih, I-Hsien Ting and Tzung-Pei Hong

8 ***K*-Neighborhood Shortest Path Privacy in the Cloud** 65
Shyue-Liang Wang, Jia-Wei Chen, I-Hsien Ting and Tzung-Pei Hong

9 **The Framework of Information Processing Network for Supply Chain Innovation in Big Data Era** 77
Chian-Hsueng Chao

10 Website Navigation Recommendation Based on Reinforcement Learning Technique 87
Yin-Ling Tang, I-Hsien Ting and Shyue-Liang Wang

11 An Approach for Hate Groups Detection in Facebook 101
I-Hsien Ting, Hsing-Miao Chi, Jyun-Sing Wu and Shyue-Liang Wang

12 Toward Crowdsourcing Data Mining 107
Hsin-Chang Yang and Chung-Hong Lee

13 Wireless Security Analysis Using WarDrive Investigation in Kaohsiung Areas 111
Hanwei Hsiao, Tienhe Chang and ChihChe Chang

14 Guanxi Buying in the Social Media Environment 123
Cathy S. Lin and Shin Yan Lu

15 Introspection of Unauthorized Sharing on Social Networking Sites 129
Cathy S. Lin and Ting-Yi Lin

Chapter 1
An Information Quality (InfoQ) Framework for Ex-Ante and Ex-Post Evaluation of Empirical Studies

Galit Shmueli and Ron Kenett

Abstract Numbers are not data and data analysis does not necessarily produce information and knowledge. Statistics, data mining, and artificial intelligence are disciplines focused on extracting knowledge from data. They provide tools for testing hypotheses, predicting new observations, quantifying population effects, and summarizing data efficiently. In these fields, measurable data is used to derive knowledge. However, a clean, exact and complete dataset, which is analyzed professionally, might contain no useful information for the problem under investigation. The term *Information Quality* (InfoQ) was coined by Ref. [15] as the potential of a dataset to achieve a specific (scientific or practical) goal using a given data analysis method. InfoQ is a function of goal, data, data analysis, and utility. Eight dimensions that relate to these components help assess InfoQ: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Generalizability, Chronology of Data and Goal, Construct Operationalization, and Communication. The eight dimensions can be used for developing streamlined evaluation metrics of InfoQ. We describe two studies where InfoQ was integrated into research methods courses, guiding students in evaluating InfoQ of prospective and retrospective studies. The results and feedback indicate the importance and usefulness of InfoQ and its eight dimensions for evaluating empirical studies.

G. Shmueli (✉)
Srini Raju Centre for IT and the Networked Economy,
Indian School of Business, Hyderabad, 500032India
e-mail: galit_shmueli@isb.edu

G. Shmueli
Rigsum Institute of IT and Management, Thimphu, Bhutan

R. Kenett
KPA Ltd., Raanana Israel Dept of Statistics & Applied Mathematics,
University of Torino, Turin, Italy

R. Kenett
Center for Finance and Risk Engineering, NYU-Poly, Brooklyn, NY 11201, USA

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_1,

Keywords Data analytics · Data mining · Statistical modeling · Study goal · Empirical study evaluation quality

1.1 Introduction and Motivation

The term Intelligent Data Analysis (IDA) implies an expectation that data analysis will yield insights and knowledge. Research and academic environments focus on developing intelligent tools for extracting information from data. Statistics education is typically aimed at teaching analysis quality. Godfrey [10] describes low quality of analysis as "poor models and poor analysis techniques, or even analyzing the data in a totally incorrect way". The book *Guide to Intelligent Data Analysis* [4] has the subtitle *How to Intelligently Make Sense of Real Data* and is focused on pitfalls that lead to wrong or insufficient analysis of results. In other words, *intelligent* most often refers to the analysis quality.

While analysis quality is critically important, another key component of IDA is the usefulness of a particular dataset for the problem at hand. The same data can contain high-quality information for one purpose and low-quality information for another purpose. An important question that arises both in scientific research and in practical applications is therefore: what is the potential of a dataset to achieve a particular goal of interest? This is related to the Zeroth Problem, coined by Mallows [18], which is the general question of "how do the data relate to the problem, and what other data might be relevant?" Hand [11] notes, "statisticians working in a research environment… may well have to explain that the data are inadequate to answer a particular question". Patzer [19] comments: "data may be of little or no value, or even negative value, if they misinform".

There is therefore a need to formalize these important aspects of IDA that have thus far not been formalized. Recently, Kenett and Shmueli coined the term *Information Quality*, or InfoQ, to define *the potential of a dataset to achieve a specific (scientific or practical) goal using a given data analysis method* ([15] with discussion and rejoinder). InfoQ lies on the interface of data, goal, and analyst and is tightly coupled with the analysis context. This is schematically illustrated in Fig. 1.1.

Fig. 1.1 InfoQ depends on data quality and analysis quality, conditional on the goal at hand

The focus of this paper is on integrating InfoQ into the thought process of data analysts, while conducting an active empirical study as well as for ex-post evaluation of empirical studies. We proceed as follows: Sect. 1.2 introduces InfoQ, its components, terminology and formal definition. In Sect. 1.3 we describe eight dimensions of InfoQ that are useful for assessing InfoQ in practice. Section 1.4 discusses an evaluation methodology, and then describes two studies. The first study describes the integration of InfoQ into a graduate-level research methods course at Ljubljana University. The second study describes an InfoQ assignment designed for ex-post evaluation of empirical studies, and its implementation in a Masters in Statistical Practice program at Carnegie Mellon University. We conclude and offer future directions in Sect. 1.5.

1.2 Information Quality: Terminology and Definition

InfoQ is a function of several components: data, analysis goal, data analysis method, and the anticipated utility from the analysis. We describe each of these four components and then define the InfoQ function.

1.2.1 InfoQ Components

Analysis Goal (g): Data analysis is used for variety of purposes. Three general classes of goals are causal explanation, prediction, and description [21, 22]. Causal explanation includes questions such as "Which factors cause the out-come?" Prediction goals include forecasting future values of a time series and predicting the output value for new observations given a set of input variables. Descriptive goals include quantifying and testing for population effects using data summaries, graphical visualizations, statistical models, and statistical tests. Deming [6] introduced the distinction between enumerative studies, aimed at answering the question "how many?" and analytic studies, aimed at answering the question why? Later, Tukey [23] proposed a classification of exploratory and confirmatory data analysis. Our use of the term goal generalizes all of these different types of goals and goal classifications.

Data (X): The term data includes any type of data to which empirical analysis can be applied. Data can arise from different collection tools: surveys, laboratory tests, field and computer experiments, simulations, web searches, observational studies and more. Data can be univariate or multivariate (one or more variables) and of any size (from a single observation in case studies to many observations). It can also contain semantic, unstructured information in the form of text or images with or without a dynamic time dimension. Data is the foundation of any application of empirical analysis.

Data Analysis Method (f): We use the term data analysis to refer to statistical analysis and data mining. This includes statistical models and methods (parametric, semiparametric, nonparametric), data mining algorithms, and graphical methods. Operations research methods, such as simplex optimization, where problems are modeled and parametrized, fall into this category as well.

Utility (U): The extent to which the analysis goal is achieved is typically measured by some performance measure. We call this measure utility. For example, in studies with a predictive goal a popular performance measure is predictive accuracy. In descriptive studies, common utility measures are goodness-of-fit measures. In explanatory models, statistical power and strength-of-fit measures are common utility measures.

1.2.2 Information Quality (InfoQ): Definition

Following Hands definition of statistics as The technology of extracting meaning from data [11], we consider the utility of applying a technology (f) to a resource (X) for a given purpose (g). In particular, we focus on the question "What is the potential of a particular dataset to achieve a particular goal using a given empirical analysis method?" To formalize this question of interest, we define the concept of Information Quality (InfoQ) as:

$$InfoQ(f, X, g) = U(f(X \mid g)) \tag{1.1}$$

InfoQ is affected by the quality of its components g (quality of goal definition), X (data quality), f (analysis quality), and U (quality of utility measure) as well as by the relationships between X, f, g and U.

1.2.3 Example: Online Auctions

Some of the large online auction websites, such as eBay, provide data on closed and ongoing auctions, triggering a growing body of research in academia and in practice. A few popular analysis goals have been:

- Determining factors affecting the final price of an auction [17]
- Predicting the final price of an auction [8]
- Descriptive characterization of bidding strategies [2, 5]
- Comparing behavioral characteristics of auction winners versus fixed-price buyers [1]
- Building descriptive statistical models of bid arrivals or bidder arrivals [5].

Given the diverse goals, it is intuitive that one dataset of eBay auctions would hold different value (InfoQ) in terms of its potential to derive insights.

Let us consider a particular goal for illustrating the components of InfoQ. Econometricians are interested in determining factors affecting the final price of an online auction. While game theory provides an underlying theoretical causal model of price in offline auctions, the online environment differs in substantial ways. We consider a study by Katkar and Reiley [13] who investigated the effect of two types of reserve prices on the final auction price. A reserve price is a value set by the seller at the start of the auction. If the final price does not exceed the reserve price, the auction does not transact. On eBay, sellers can choose to place a public reserve price that is visible to bidders, or an invisible secret reserve price (bidders only see that there is a reserve price but do not know its value). InfoQ, in the context of this study, consists of asking the question: "Given the data collected on a set of auctions, what is their potential to allow quantifying the difference between secret and public reserve prices using regression modeling?"

Study Goal (g):	Quantify the effect of using a secret versus public reserve price on the final price of an auction.
Data (X):	The authors conducted a field experiment by selling Pokemon cards on eBay. They auctioned 25 identical pairs of Pokemon cards in week-long auctions during a 2-week period in April 2000, where each card was auctioned twice: once with a public reserve price and once with a secret reserve price. The resulting data included the complete information on all 50 auctions.
Data Analysis Method (f):	The authors used linear regression to test for the effect of private/public reserve price on the final auction price and to quantify it.
Utility (U):	The authors used statistical significance (p value) of the regression coefficient to assess the presence of an effect for private/public reserve price. They used the regression coefficient value for quantifying the magnitude of the effect (they conclude: "a secret-reserve auction will generate a price $0.63 lower, on average, than will a public-reserve auction".)

1.3 Eight Dimensions of InfoQ

Quality of Statistical Data is a concept developed and used in European official statistics and international organizations such as the International Monetary Fund (IMF) and The Organisation for Economic Cooperation and Development (OECD). This concept refers to the usefulness of summary statistics produced by national statistics agencies and other producers of official statistics. This is a special case of InfoQ, where the data analysis method (f) is the computation of summary statistics.

Although this operation might seem very simple, it is nonetheless considered analysis, because it is in fact estimation. Hence, InfoQ is more general. Quality of statistical data is evaluated in terms of the usefulness of the statistics for a particular goal. The OECD uses seven dimensions for quality assessment: relevance, accuracy, timeliness and punctuality, accessability, interpretability, coherence, and credibility (Chap. 5 in [9]). We use a similar framework to determine InfoQ dimensions.

Taking an approach similar to data quality assessment, we define eight dimensions for assessing InfoQ that consider and affect not only the data and goal, but also the analysis method and utility function. With this approach we provide a decomposition of InfoQ that can be used for assessing and improving research initiatives or ex-post evaluations.

1.3.1 Data Resolution

Data resolution refers to the measurement scale and aggregation level of X. The measurement scale of the data should be carefully evaluated in terms of its suitability to the goal, the analysis methods to be used, and the required resolution of U. Given the original recorded scale, the researcher should evaluate its adequacy. It is usually easy to produce a more aggregated scale (e.g., two income categories instead of ten), but not a finer scale. Data might be recorded by multiple instruments or by multiple sources. To choose among the multiple measurements, supplemental information about the reliability and precision of the measuring devices or data sources is useful. A finer measurement scale is often associated with more noise; hence the choice of scale can affect the empirical analysis directly. The data aggregation level must also be evaluated relative to the goal.

1.3.2 Data Structure

Data structure relates to the type(s) of data and data characteristics such as corrupted and missing values due to the study design or data collection mechanism. Data types include structured numerical data in different forms (e.g., cross-sectional, time series, network data) as well as unstructured, non-numerical data (e.g., text, text with hyperlinks, audio, video, and semantic data). The InfoQ level of a certain data type depends on the goal at hand.

1.3.3 Data Integration

With the variety of data source and data types, there is often a need to integrate multiple sources and/or types. Often, the integration of multiple data types creates

new knowledge regarding the goal at hand, thereby increasing InfoQ. For example, in online auction research, the integration of temporal bid sequences with cross-sectional auction and seller information has led to more precise predictions of final prices (see Chap. 4 in [12]) as well as to an ability to quantify the effects of different factors on the price process [3].

1.3.4 Temporal Relevance

The process of deriving knowledge from data can be put on a time line that includes the data collection, data analysis, and study deployment periods as well as the temporal gaps between the data collection, the data analysis, and the study deployment stages (see Fig. 1.2). These different durations and gaps can each affect InfoQ. The data collection duration can increase or decrease InfoQ, depending on the study goal, e.g., studying longitudinal effects versus a cross-sectional goal. Similarly, if the collection period includes uncontrollable transitions, this can be useful or disruptive, depending on the study goal.

1.3.5 Chronology of Data and Goal

The choice of variables to collect, the temporal relationship between them, and their meaning in the context of the goal at hand also affects InfoQ. For example, in the context of online auctions, classic auction theory dictates that the number of bidders is an important driver of auction price. Models based on this theory are useful for explaining the effect of the number of bidders on price. However, for the purpose of predicting the price of ongoing online auctions, where the number of bidders is unknown until the auction end, the variable "number of bidders", even if available in the data, is useless. Hence, the level of InfoQ contained in number of bidders for models of auction price depends on the goal at hand.

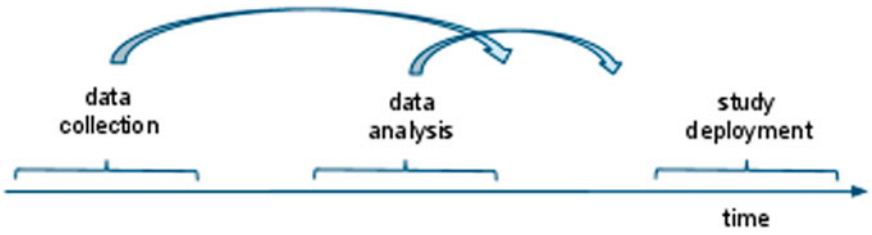


Fig. 1.2 Temporal durations and gaps that affect InfoQ. Feedback *arrows* indicate the cyclic process of data collection and analysis

1.3.6 Generalizability

The utility of $f(X \mid g)$ is dependent on the ability to generalize f to the appropriate population. Two types of generalizability are statistical and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability refers to applying a model based on a particular target population to other populations. This can mean either generalizing an estimated population pattern/model f to other populations, or applying f estimated from one population to predict individual observations in other populations. Determining the level of generalizability requires careful characterization of g. For instance, for inferring about a population parameter, statistical generalizability and sampling bias are the focus, and the question of interest is "What population does the sample represent?". In contrast, for predicting the values of new observations, the question of interest is whether f captures associations in the training data X that are generalizable to the to-be-predicted data.

1.3.7 Construct Operationalization

Constructs are abstractions that describe a phenomenon of theoretical interest. Measurable data are an operationalization of underlying constructs. The relationship between the underlying construct and its operationalization can vary, and its level relative to the goal is another important aspect of InfoQ. The role of construct operationalization is dependent on the goal, and especially on whether the goal is explanatory, predictive, or descriptive. In explanatory models, based on underlying causal theories, multiple operationalizations might be acceptable for representing the construct of interest. As long as the data are assumed to measure the construct, the variable is considered adequate. In contrast, in a predictive task, where the goal is to create sufficiently accurate predictions of a certain measurable variable, the choice of operationalized variable is critical.

1.3.8 Communication

Effective communication of the analysis and its utility directly impacts InfoQ. There are plenty of examples where miscommunication of valid results has led to disasters, such as the NASA shuttle Challenger disaster [16]. Communication media are visual, textual, and verbal presentations and reports. Within research environments, communication focuses on written publications and conference presentations. Research mentoring and the refereeing process are aimed at improving communication and InfoQ within the research community.

1.4 Assessing Information Quality

The eight dimensions of InfoQ are intended to help operationalize the concept into actionable evaluation. Considering InfoQ and its dimensions can be useful during an ongoing empirical study (from study design, through data collection and analysis, to presentation) as well as in ex-post evaluations of completed studies. InfoQ evaluation can range from qualitative to quantitative. Qualitative evaluation consists of verbal or written assessments of a study on each dimension. Quantitative evaluation requires defining a metric for which we can evaluate each dimension. One such rating-based metric is described in the next section.

1.4.1 Rating-Based Evaluation

Similar to the use of "data quality" dimensions by statistical agencies for evaluating data quality, we evaluate the eight InfoQ dimensions to assess InfoQ. This evaluation integrates different aspects of a study and assigns an overall InfoQ score. The broad perspective of InfoQ dimensions is designed to help researchers enhance the added value of their studies.

Assessing InfoQ using quantitative metrics can be done in several ways. Reference [15] presented a rating-based approach that examines a study report and scores each of the eight InfoQ dimensions. A coarse grained approach is to rate each dimension on a 1–5 scale, with "5" indicating "high" achievement on that dimension. The ratings (Y_i, i = 1, …, 8) can then be normalized into a desirability function (see [7]) for each dimension, ($0 \leq d(Y_i) \leq 1$), which are then combined to produce an overall InfoQ score using the geometric mean of the individual desirabilities:

$$\text{InfoQ Score} = [d_1(Y_1) \times d_2(Y_2) \times \ldots \times d_8(Y_8)]^{1/8} \tag{1.2}$$

In the next sections we describe two studies where participants used InfoQ dimensions to evaluate an empirical study—either a proposed study or a completed one. The goal of the two studies was to introduce graduate students to the application of statistics in practice, to key questions that a statistician should ask, and to the link between goal, data, analysis and context. The InfoQ framework provides such an integrative view.

1.4.2 Study 1: Evaluating Research Proposals

Graduate students at the Faculty of Economics of the University of Ljubljana undergo a 3-day workshop on research methods, to equip them with methodology for developing and presenting a research proposal that is the basis for their

doctorate thesis. One of the first milestones for students is to defend their research proposal in front of a committee.

The class consisted of approximately 50 graduate students from a wide range of areas, including organizational behavior, operations research, marketing, and economics. In 2009, InfoQ was integrated into the research methods workshop. The goal was to help students (and their advisors) figure out whether their proposed research is properly defines as to potentially generate effective knowledge.

Students worked in small teams and discussed the InfoQ dimensions of their draft proposal. Each student then gave a 15 min presentation to the whole team. Details of the workshop and pre-workshop assignments are available at goo.gl/f6bIA. Students' grades in the workshop were derived from an InfoQ score of their proposal submission, which consisted of a PowerPoint presentation and a written document. This approach was designed to make their research journey more efficient and more effective. Feedback by students and faculty, has indicated that the InfoQ-based research methods workshop has indeed met this goal [14].

1.4.3 Study 2: Ex-Post Evaluation of Empirical Studies

We designed an assignment that requires participants to evaluate five empirical studies based on written reports. Based on the reports and on a brief introduction to InfoQ, participants were asked to:

(1) give a brief description of the goal, data, analysis, and utility measure for each study, and
(2) rate the study on each of the eight InfoQ dimensions

The form with information on InfoQ, on the five studies, and the InfoQ questions and ratings are available at goo.gl/erNPF. Figure 1.3 shows the questions asked for one of the studies.

In 2012, the InfoQ assignment was integrated into a course in the Masters in Statistics Practice program at Carnegie Mellon University's Statistics Department. Each of the 16 students spent about 60–90 min reading the 5 studies and evaluating the 8 dimensions for each study.

Comparing the responses of the 16 participants on each of the five studies revealed variability in respondents ratings of the InfoQ dimensions. This variability indicates a need to further streamline the process of quantifying each dimension rating. An important result of this study was the feedback regarding the value added by going through the evaluation process. Participants reported that using this approach helped them "sort out all of the information", and several reported that they will adopt this evaluation approach for future studies.

Evaluating Information Quality (InfoQ)

Study #1: Predicting days with unhealthy air quality in Washington DC
Several tour companies' revenues depend heavily on favorable weather conditions. This study looks at air quality advisories, during which people are advised to stay indoors, within the context of a tour company in Washington DC.

You will find the study report and presentation here (copy and paste the address into your browser): http://galitshmueli.com/content/tourism-insurance-predicting-days-unhealthy-air-quality-washington-dc

Stated objective of study
Is the stated objective to explain, predict, or to describe? If more than one objective is stated, choose the highest priority objective

- Explain (how or which factors affect air quality)
- Predict (air quality on new or future days)
- Describe (the relationship between air quality and other variables)

Data used and its origin
Briefly describe the data used in the analysis

Analysis methods used in study
List all the methods mentioned (for example: histogram, logistic regression)

Utility of findings
What is the potential value of the study findings? To whom? When?

Kindly rate the project on the 8 InfoQ dimensions
Information about the 8 dimensions is available at http://tinyurl.com/6pmrwcx

	completely inadequate	inadequate	reasonable	achieved	fully achieved
Data resolution					
Data structure					
Data integration					
Temporal relevance					
Generalizability					
Chronology of data and goal					
Construct operationalization					
Communication					

Additional comments
Feel free to write any additional comments that you may have regarding this study

Fig. 1.3 InfoQ evaluation form for an empirical study on air quality. The complete form with information and additional studies for evaluation are available at goo.gl/erNPF

1.5 Conclusions and Future Directions

Our discussion of InfoQ as a crucial component in the empirical analysis framework calls for further discussion and research in various directions. In assessing InfoQ, we proposed a rating-based approach. We describe two initial studies that attempt to gauge the effectiveness of using the InfoQ framework for developing an analysis plan and for evaluating an empirical study. Future research is needed on specific implementations such as investigating the reliability of ratings across raters.

Although we discussed several dimensions of InfoQ and their relation to other quality concepts, there exist others that might be considered and new dimensions might evolve over time. For example, in todays environment, an important aspect of InfoQ is related to data privacy and confidentiality. In some areas, InfoQ could include a measure of risk in terms of confidentiality or human subjects. We also note the relationship with Big Data dimensions: the first five InfoQ dimensions relate to the three V's of Big Data [20]: *Volume* (data resolution; data structure), *Velocity* (temporal relevance; chronology of data and goal), *Variety* (data structure; data integration). Other proposed V's (value, veracity, viscosity, virality) can also be related to InfoQ dimensions, depending on their definitions.

Our studies have indicated the usefulness of considering InfoQ and its eight dimensions for evaluating prospective and retrospective studies. The InfoQ framework helps formalize and streamline the informal process that an experienced data analyst goes through. Future work will focus on streamlining rating-based and other InfoQ evaluation methods.

Acknowledgments We thank Professors Joel Greenhouse (Carnegie Mellon University), Shirley Coleman (Newcastle University), and Irena Ograjenek (University of Ljubljana) for their support of integrating InfoQ into graduate courses at CMU and University of Ljubljana, and helping assess its impact.

References

1. Angst CM, Agarwal R, Kuruzovich J (2008) Bid or buy? Individual shopping traits as predictors of strategic exit in on-line auctions. Int J Electron Commer 13:59–84
2. Bapna R, Goes P, Gupta A, Jin Y (2004) User heterogeneity and its impact on electronic auction market design: an empirical exploration. MIS Quarterly, 28(1):21
3. Bapna R, Jank W, Shmueli G (2008) Price formation and its dynamics in online auctions. Decis Support Syst 44:641–656
4. Berthold MR, Borgelt C, Hoppner F, Klawonn F (2010) Guide to intelligent data analysis. Springer, London
5. Borle S, Boatwright P, Kadane JB (2006) The timing of bid placement and extent of multiple bidding: an empirical investigation using eBay online auctions. Stat Sci 21:194–205
6. Deming WE (1953) On the distinction between enumerative and analytic studies. J Am Stat Assoc 48:244–255

7. Figini S, Kenett RS, Salini S (2010) Integrating operational and financial risk assessments. Qual Reliab Eng Int 26(8):887–897
8. Ghani R, Simmons H (2004) Predicting the end-price of online auctions. Pisa, Italy
9. Giovanni E (2008) Understanding economic statistics. Technical report
10. Godfrey AB (2008) Eye on data quality. *Six Sigma Forum Magazine*, pp 5–6
11. Hand DJ (2008) Statistics: a very short introduction. Oxford University Press, Oxford
12. Jank W, Shmueli G (2010) Modeling online auctions. Wiley, Hoboken
13. Katkar R, Reiley DH (2006) Public versus secret reserve prices in eBay auctions: results from a pokémon field experiment. *Advances in Econc Analysis and Policy*, 6(2), Article 7, 1–23
14. Kenett RS, Coleman S, OgrajenÅ¡ek I (2010) On quality research: an application of InfoQ to the Phd research process. In: Proceedings of the European network for business and industrial statistics (ENBIS) 10th annual conference on business and industrial statistics, Antwerp, Belgium, September 2010
15. Kenett RS, Shmueli G (2013) On information quality. J Roy Stat Soc Ser A, forthcoming
16. Kenett RS, Thyregod P (2006) Aspects of statistical consulting not taught by academia. Stat Neerl 60:396–412
17. Lucking-Reiley D, Bryan D, Prasad N, Reeves D (2007) Pennies from eBay: the determinants of price in online auctions. J Ind Econ 55:223–233
18. Mallows C (1998) The zeroth problem. Am Stat 52:1–9
19. Patzer GL (2005) Using secondary data in marketing research. Praeger, Westport
20. Russom P (2011) Big data analytics. Technical report, Q4
21. Shmueli G (2010) To explain or to predict? Stat Sci 25:289–310
22. Shmueli G, Koppius OR (2011) Predictive analytics in information systems research. Manag Inf Syst Q 35:553–572
23. Tukey JW (1977) Exploratory data analysis. Addison Wesley, New York

Chapter 2
Memory-Aware Mining of Indirect Associations Over Data Streams

Wen-Yang Lin, Shun-Fa Yang and Tzung-Pei Hong

Abstract In this study, we focus on over a data stream the mining of indirect associations, a type of infrequent patterns that reveal infrequent itempairs yet highly co-occurring with a frequent itemset called "mediator". We propose a generic framework MA-GIAMS, an extension of the GIAMS framework with memory-awareness capability that can cope with the variation of available memory space, making use of most available memory to accomplish the discovery of indirect association rules without incurring too much overhead and retaining as could as possible the accuracy of discovered rules. Empirical evaluations show that our algorithm can efficiently adjust the size of the data structure without sacrificing too much the accuracy of discovered indirect association rules.

Keywords Adaptation scheme · Indirect association · Memory constraint · Resource-awareness · Stream mining

2.1 Introduction

As the advent of emerging techniques in the information explosion age, e.g., internet, handheld devices, RFID, sensor network, e-commerce, etc., more and more systems or applications would generate continuous rapid flow of vast data in a timely and endless fashion. This heralds a new type of data source, called data streams, and brings new challenges to the data mining research community. Unlike

W.-Y. Lin (✉) · S.-F. Yang · T.-P. Hong
Department of Computer Science and Information Engineering,
National University of Kaohsiung, Kaohsiung, Taiwan
e-mail: wylin@nuk.edu.tw

T.-P. Hong
e-mail: tphong@nuk.edu.tw

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_2,

traditional static data sources, streaming data is fast changing, continuously generated and unbounded in amount. It is nearly impossible to store a stream entirely in a persistent storage, making contemporary mining algorithms designed for static dataset awkward and inapplicable. Recent studies on stream data mining have achieved some general requirements for effective mining algorithms [4], including single-pass of scanning over data set, real-time execution, and low memory consumption.

Not until recently, however, researchers have begun to pay attention to another critical and more challenging issue, adaptive mining of streaming data with respect to variations in available resources, e.g., CPU power and memory space. Any mining algorithm running on these resources constrained devices has to be able to adjust and adapt its execution to utilize the very limited, changing in availability resources. Contemporary research work on resource-aware mining over data streams can be viewed from two aspects, the type of resources under concern, e.g., CPU power [3] or memory space [5], and the type of mining tasks employed, e.g., clustering [10], density estimation [8], frequent itemset mining [3, 13], etc.

From the viewpoint of data mining task, almost all works were focusing on frequent pattern discovering; no work, to our knowledge, has been devoted to the problem of mining infrequent patterns, e.g., indirect associations. Another minor restriction of the previous works is that they were conducted with respect to some specific stream window model, e.g., landmark, time-fading, or sliding window. Algorithms tailored for some particular window model usually are not applicable to and so need redesigning to fit to other models.

In this study, we aim at developing a generic mining framework with memory-aware capability that can adapt the computation in accordance with data arriving rate as well as the available memory space. Specifically, we focus on the mining of indirect associations [11], a type of infrequent patterns that reveal infrequent itempairs yet highly co-occurring with a frequent itemset called "mediator". For example, Coca-cola and Pepsi are competitive products and could be replaced by each other. So it is very likely that there is an indirect association rule revealing that consumers buy a kind of cookie tend to buy together with either Coca-cola or Pepsi but not both, denoted as ⟨Coca-cola, Pepsi | cookie⟩. Below is a formal definition of indirect association.

Definition 1 An itempair {a, b} is indirectly associated via a mediator M, denoted as ⟨a, b | M⟩ if the following conditions hold:

1. $\sup(\{a, b\}) < \sigma_s$ (Itempair support condition);
2. $\sup(\{a\} \cup M) \geq \sigma_f$ and $\sup(\{b\} \cup M) \geq \sigma_f$ (Mediator support condition);
3. $dep(\{a\}, M) \geq \sigma_d$ and $dep(\{b\}, M) \geq \sigma_d$ (Mediator dependence condition); where sup(A) denotes the support of an itemset A, and dep(P, Q) is a measure of the dependence between itemsets P and Q.

We propose a generic framework MA-GIAMS with memory-awareness capability, an extension of our previously proposed GIAMS framework [7] that can represent all classical streaming models and retain user flexibility in defining new

models. Our MA-GIAMS framework can cope with the variation of available memory space, making use of most available space to accomplish the discovery of indirect association rules without too much overhead and retaining as could as possible the accuracy of discovered rules. To realize the memory awareness scheme, we propose a victim searching and node releasing algorithm to adjust the structure for maintaining potential frequent itemsets in accordance with the available memory space. Empirical evaluations on a real dataset show our algorithm can efficiently adjust the size of the structure without sacrificing too much the accuracy of discovered indirect association rules.

The remainder of this paper is organized as follows. In Sect. 2.2, we provide some background knowledge and related work about resource-aware stream mining. Since our work in this study is based on the GIAMS framework, we give an overview of GIAMS in Sect. 2.3. Section 2.4 describes our proposed MA-GIAMS framework, including the algorithmic design for adaptive functionalities with respect to memory space variation, and the detailed data structure and procedure for realizing the algorithmic framework. Section 2.5 describes the experiments we conducted on evaluating the proposed framework. Finally, the conclusions and future works are presented in Sect. 2.6.

2.2 Related Work

The research of resource-aware stream mining focuses mainly on how to use the limited resources to efficiently accomplish the mining task while guarantee as could as possible the accuracy of mining results. Contemporary research work on resource-aware mining over data streams can be viewed from three aspects, the type of resources under concern, the type of mining tasks employed, and the type of adaptation techniques used.

Studies conducted from the viewpoint of resource type consider CPU power and memory space, or additional issues for mobile devices, battery and network bandwidth. The types of mining tasks studied include clustering, frequent pattern mining, kernel density estimation. Finally, the adaptation techniques can be input adaptation or algorithm adaptation. The input adaptation technique refers to schemes used in adjusting the amount of input data to keep up with the pace of the stream and meet the computing capacity. Three commonly used approaches including sampling, i.e., statistically choosing some input data, load shedding, i.e., discarding part of the input data, and data synopsis creation, referring to data summarization techniques that retain the data characteristics, statistics or profile.

Teng et al. [13] proposed a wavelet based data synopsis technique, called Resource-Aware Mining for Data Streams (RAM-DS), to transform the input data stream into different granularity of data from temporal view or frequency view in order to reduce the data amount. Their work focused on frequent pattern mining.

The group led by M.M. Gaber is one of the pioneers in resource-aware stream data mining. They have conducted a series of research studies on resource-

awarestream clustering [5, 6], with ultimate goal at developing a general resource-aware framework that can adapt to variability of different resource availability over time. Their recent work [11] proposed a generic framework called Algorithm Granularity Settings, which uses three levels of adaptation strategies in the data input, algorithm processing, and data output, and cope with different issues of resource-awareness.

The work conducted by Dang et al. [2, 3] considers stream mining under CPU resource constraint, focusing on frequent pattern discovery, and using load shedding technique. Their approach relies on a way to estimate the system workload by approximately computing the number of maximal itemsets that can be generated from a transaction, and then employs the load shedding technique to sample transactions if the system workload is over the current CPU computing power.

Heinz and Seeger [8] proposed an algorithm adaptation method that is tailored to the problem of kernel density estimation. The resource type considered in their work is memory space.

In distributed computing environment, the network bandwidth available for data transmission is particularly important. Parthasarathy and Subramonian [9] developed a resource-aware scheduling scheme to cope with the bandwidth limitation.

2.3 Overview of GIAMS

Generic Indirect Association Mining over Streams (GIAMS) [7] is an algorithmic framework that can accommodates the three stream window model called Landmark model, Time-fading model and Sliding window, while retains user flexibility for defining new models. Users only have to set four variables (timestamp, window size, stride and decay rate) in accordance with the generic window model, then GIAMS can discover indirect association rules.

Suppose that we have a data stream $S = (t_0, t_1, t_{2,\ldots}t_i,\ldots)$, where t_i denotes the transaction arrived at time i. Since data stream is a continuous and unlimited incoming data along with time, a window W usually is specified, representing the sequence of data arrived from t_i to t_j, denoted as $W[i, j] = (t_i, t_{i+1},\ldots, t_j)$. GIAMS adopts a generic window model Ψ for data stream mining, which is dictated as a four-tuple specification, $\Psi(l, w, s, d)$, where l denotes the timestamp at which the window start, w as the window size, s is the stride the window moves forward, and d is the decay rate. The stride notation s is introduced to allow the window moving forward in a batch (block) of transactions (of size s). In this regard, the stride notation s also specifies the size of block.

The GIAMS framework is developed according to the paradigm proposed by Tan et al. [12]: First, discovers the set of frequent itemsets with support higher than σ_f, and then generates the set of qualified indirect associations from the frequent itemsets. Based on this paradigm, GIAMS works in the following scenario: (1) The user first sets the streaming window model by specifying the parameters described

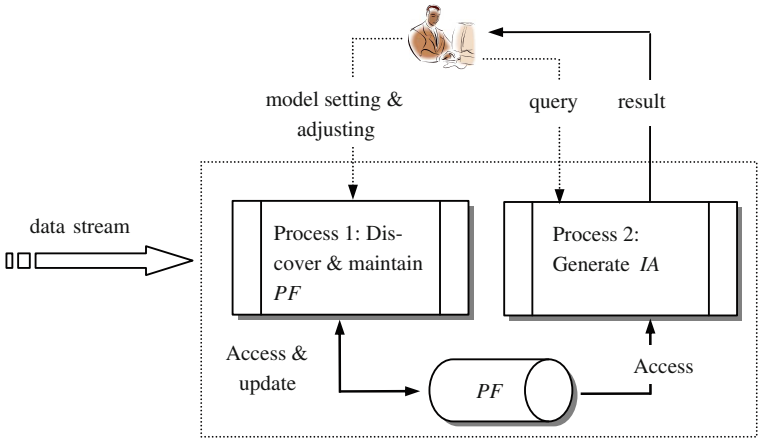


Fig. 2.1 The GIAMS generic framework for indirect association mining

previously; (2) The framework then executes the process for discovering and maintaining the set of potential frequent itemsets *PF* as the data continuously stream in; and (3) At any moment once the user issues a query about the current indirect associations the second process for generating the qualified indirect associations is executed to generate from *PF* the set of indirect associations *IA*. Figure 2.1 depicts the generic streaming framework for indirect associations mining.

2.4 Adaptation Scheme for Available Memory Awareness

2.4.1 Data Structure

The structure used in our MA-GIAMS for realizing *PF* is a modification of the tree structure used in GIAMS, called *Card-Stree*, which is a forest of search trees keeping itemsets of different cardinalities, appearing in the current window, say ST_1, ST_2, …, ST_k, for ST_k maintaining the set of frequent k-itemsets. We name the modified structure *Card-Stree**. Each node in *Card-Stree** except the root keep the information of the maintained itemset. More specifically, for each itemset X, the node records *X.id*, the identifier of X; *X.bidv*, the vector of identifiers of the blocks that X appears; *X.countv*, the vector that stores the number of occurrences of X within each block; and *X.tlcount*, the total occurrences that X appears in the current window. An example of *Card-Stree** after processing the example data stream shown in Fig. 2.2 is illustrated in Fig. 2.3. Because each node consumes approximately the same amount of memory space, in what follows we use a node as the memory unit.

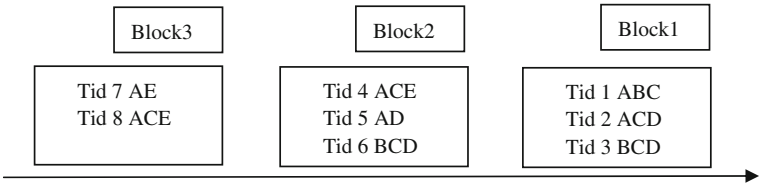


Fig. 2.2 An example data stream

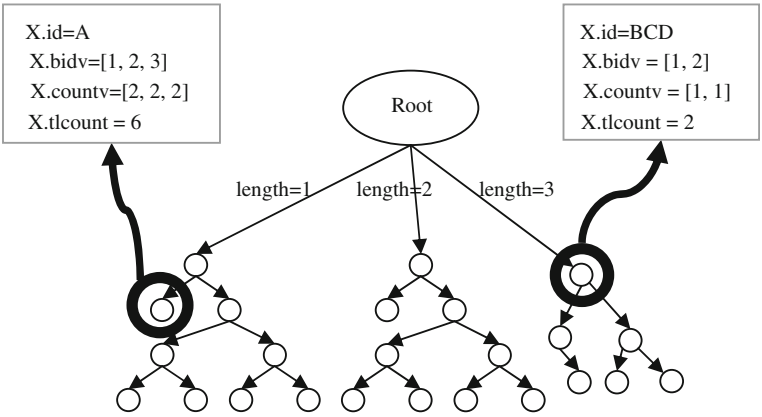


Fig. 2.3 An illustration of *Card-Stree**

2.4.2 Adaptive Node Releasing

A simple and intuitive approach is blindly dropping some itemsets while the memory space is not enough. This approach, however, would loss too much information, making the mining results incorrect and leading to wrong analysis. Rather, we employ a strategy similar to the concept of cache replacement. When the memory space is insufficient, we decide which itemsets in the current *Card-Stree** are less important and can be deleted to release enough space for accommodating the incoming, more important itemsets.

In this regard, we propose a node releasing mechanism to cope with the situation when memory space is not enough to maintain all of potential frequent itemsets in the *Card-tree**. Note that the processing of stream mining needs to be computed in real time. As such, the main design concern is the efficiency, i.e., how to efficiently search and determine the victim nodes for deletion, without sacrificing too much the accuracy of the discovered rules.

Firstly, we note that for mining indirect associations, the set of 2-itemsets is the most important set, because from which all length-2 mediators and the infrequent itempairs are generated. Our node replacement thus is executed only when the memory space is not enough and the new generated itemsets from the incoming

transaction are of lengths 1 and 2. In other words, those new generated k-itemsets with $k > 2$ are discarded immediately when the memory is not enough.

Secondly, since the frequency of long itemsets is usually less than that of short itemsets, and lengthy rules constructed from long itemsets are less understandable to the users, our approach replaces nodes according to their cardinalities, first choosing the longest itemsets. More precisely, suppose that we need to release n nodes and let k denote the largest length of itemsets in *Card-Stree**. Our approach will search for the top n nodes with the smallest counts in the ST_k subtree. If there are less than n nodes found in ST_k, then the search continues in subtrees ST_{k-1}, ST_{k-2}, and so on. However, during the search process, we will delete any node whose count is equal to 1 and decrement the number of nodes to be released. This is because these nodes represent the least occurring itemsets.

In addition, to speedup the search of n victim nodes, we introduce a link structure called *victim-list* to maintain the nodes in *Card-Stree** chosen for deletion. Each node in *victim-list* contains three fields, the count of the itemsets, the number of itemsets having this count, and pointers to all the corresponding nodes in the *Card-Stree**. All nodes in *victim-list* are sorted in decreasing order of counts.

There are some remarks noticeable in the aforementioned procedure. First, the victims are maintained and deleted in groups, distinguishing by count. That is why we may end up with more than n victims in *victim-list*. Second, for efficiency concern, we only compare the count of the node X under concern with that of the first group in *victim-list*. This avoids the overhead for linear searching along the entire *victim-list*.

2.4.3 Algorithm Description

We only brief the modification to the Process 1 in GIAMS as shown in Fig. 2.1; there is no change to Process 2. In summary, for each k-itemset X generated from the input block of transactions, if X is already in *Card-Stree**, we update its information. Otherwise, if the memory space is enough, then insert X into *Card-Stree**. On the other hand, if the memory is not enough and $k \geq 3$, then X is discarded. But for $k < 3$, we temporarily store X into a buffer. After all itemsets generated from the current transaction have been inspected, Process 1 will call algorithm Victim_Searching&Releasing described in Sect 2.4.2 to release memory, and insert at most #*victim* of the new generated k-itemsets in buffer into *Card-Stree**.

2.5 Experimental Results

To evaluate the effectiveness and efficiency of MA-GIAMS, we conducted a preliminary experiment on a real dataset msnbc [1], which was constructed from the web log of news pages in msn.com for the entire day of September, 28, 1999,

Table 2.1 Parameter settings for generic window model used in this experiment

s	w	d	σ_s	σ_f	σ_d
10,000	80,000	1	0.01	0.01	0.1

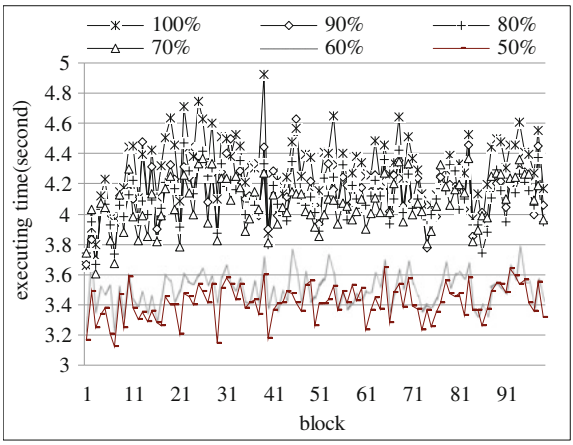


Fig. 2.4 Execution times of MA-GIAMS running over msnbc with available memory variation

which consists of 989818 transactions. More detailed description of this dataset can be found in [1]. The evaluation was inspected from two aspects, execution time and pattern accuracy. In this evaluation, we consider the sliding window model, with detailed settings of the generic model shown in Table 2.1. Since $s = 10{,}000$, the test set was divided into 99 blocks. All experiments were done on Intel(R) Core(TM) i5-2400(3.1G) PC with 4 GB of main memory, running the Windows 7 32-bit operation system. All programs were implemented in Visual C++ 2008.

To inspect the performance and effectiveness of our memory awareness adaptation scheme, we run MA-GIAMS under five settings of available memory, i.e., ranging from 90 to 50 % of the original memory space, with 10 % decrement. And compare the results with those running with sufficient memory space.

First, we evaluate the execution times of MA-GIAMS. The results are depicted in Fig. 2.4, where x-axis denotes the block number. As the results demonstrate, most of the time MA-GIAMS is faster than GIAMS even MA-GIAMS incurs overhead for running victim searching and node releasing to cope with insufficient memory. And the execution times increase as the available memory increase. This is because when the memory is not sufficient there are certain amount of itemsets that originally have to be maintained in the *Card-Stree** are pruned, which has the effect in decreasing the number of create and update operations, without doing too much of node replacement operations.

Next, we inspect the accuracy of the results generated by MA-GIAMS. Because our algorithm introduces storage shedder to prune itemsets to be maintained in the *Card-Stree** structure, so error may occur to the discovered frequent itemsets,

including the missing rate and support missing rate of frequent itemsets, and the precision and recall with respect to indirect association rules.

We first evaluate how much of frequent itemsets generated by GIAMS without memory limitation will be lost when memory shortage occurs, which is measured as error rate:

$$Miss = |F_{\text{true}} \cap F_{\text{est}}| / |F_{\text{true}}| \tag{2.1}$$

where F_{true} denotes the set of frequent itemsets discovered by GIAMS while F_{est} represents that by MA-GIAMS with insufficient memory. We also check the difference between the supports of the discovered frequent itemsets with and without memory limitation, which is measured by the following formula called Average Support Error (ASE):

$$ASE = \frac{\sum_{x \in F} (Tsup(x) - Esup(x))}{|F|} \tag{2.2}$$

where *Tsup* denotes the frequent itemsets that are discovered by GIAMS, *Esup* denotes the frequent itemsets discovered by MA-GIAMS under memory limitation. As the results displayed in Fig. 2.5, the missing rate of frequent itemsets generated by our MA-GIMAS are nearly zero for sufficient memory and are less than 0.2 even the available memory is very restricted, e.g., 60 and 50 %. All ASEs are nearly zero in all cases, even in the case that only 50 % of memory is available.

We then examine the accuracy of rules discovered by MA-GIAMS. We consider two measurements, *precision* and *recall*. Let IA_{true} denote the set of indirect associations discovered by GIAMS without memory limitation and IA_{est} denote the set discovered by MA-GIAMS with insufficient memory. The precision measures the ratio of how many indirect associations in IA_{est} are also in IA_{true}, while recall examines the percentage of how many indirect associations in IA_{true} are missed generated by MA-GIAMS. These two criteria are define as follows:

$$Precision = |IA_{\text{true}} \cap IA_{\text{est}}| / |IA_{\text{est}}| \tag{2.3}$$

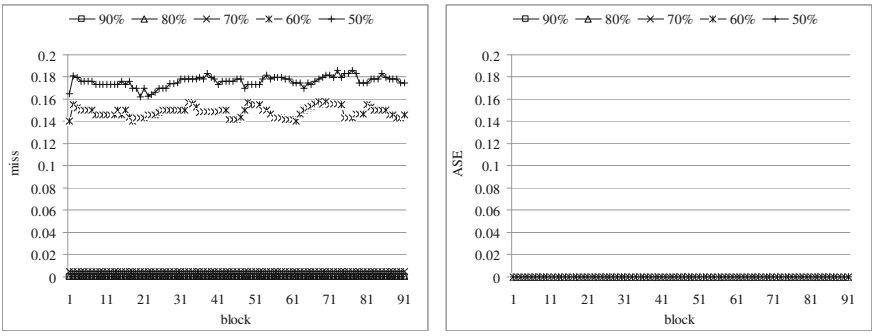


Fig. 2.5 Miss and ASE of MA-GIAMS over msnbc with available memory variation

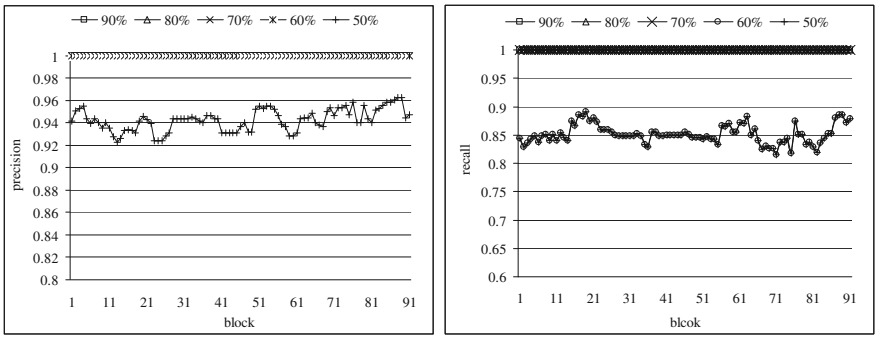


Fig. 2.6 Precisions and recalls of MA-GIAMS in terms of discovered indirect associations from msnbc with available memory variation

$$Recall = |IA_{\text{true}} \cap IA_{\text{est}}| / |IA_{\text{true}}| \tag{2.4}$$

As the results illustrated in Fig. 2.6, the memory adaptation scheme of our MA-GIAMS performs very well. All of the precisions are larger than 0.9 and recalls are above 0.8, meaning the percentages of false indirect associations discovered by our MA-GIAMS are less than 10 % and the percentages of true indirect associations not discovered by MA-GIAMS are less than 20 %, respectively, even in the case that only 50 % of memory is available.

2.6 Conclusions

In this paper, we have considered the problem of memory-aware mining of indirect association rules over data streams. We have proposed a generic framework MA-GIAMS to cope with this problem. Our proposed framework is based on GIAMS, a mining framework that can accommodate most of the contemporary stream window models and allow user-defined specific window models. Our framework add some mechanisms to GIAMS, including a resource monitor, a load shedder, and a storage shedder, as a whole can adapt the computation in accordance with data arriving rate as well as the available memory space. We have conducted a preliminary experiment to evaluate the effectiveness and performance of the proposed framework. The experimental results showed that our framework can effectively adjust the memory consumption during the course of frequent pattern mining with very little overhead. The results also showed that even with limited memory space, i.e., only 50 % of the original space, our framework not only can discover most of the frequent patterns serving as mediators for qualified indirect associations but also maintain the accuracy of discovered patterns. More experiments on over types of real datasets as well as synthetic datasets will be performed in the near future.

The study of resource-aware mining from streaming data is still in its infancy. Many research issues are worthy of further investigation. In this study, we confine the resource to memory space. Many applications developed in mobile environment, such as sensor networks, intelligent cell phones, however, have to consider additional resources constraint, mainly battery and network bandwidth. We will extend our framework to accommodate these new types of resources.

Acknowledgments This work is partially supported by National Science Council of Taiwan under grant No. NSC97-2221-E-390-016-MY2.

References

1. Cadez I, Heckerman D, Meek C, Smyth P, White S (2000) Visualization of navigation patterns on a web site using model-based clustering. In: 6th ACM SIGKDD international conference on knowledge discovery and data mining, pp 280–284
2. Dang XH, Ng WK, Ong KL (2006) Adaptive load shedding for mining frequent patterns from data streams. In: International conference on data warehousing and knowledge discovery, pp 342–351
3. Dang XH, Ng WK, Ong KL, Lee VCS (2007) Discovering frequent sets from data streams with CPU constraint. In: 6th Australasian data mining conference, pp 121–128
4. Domingos P, Hulten G (2003) A general framework for mining massive data streams. J Comp Graph Stat 12(4):945–949
5. Gaber MM, Krishnaswamy S, Zaslavsky A (2005) Resource-aware mining of data streams. J Univers Comp Sci 11(8):1440–1453
6. Gaber MM, Yu PS (2006) A framework for resource-aware knowledge discovery in data streams: a holistic approach with its application to clustering. In: ACM symposium on applied computing, pp 649–656
7. Lin WY, Wei YE, Chen CH (2011) A generic approach for mining indirect association rules in data streams. In: 24th international conference on industrial, engineering and other applications of applied intelligent systems, pp 95–104
8. Heinz C, Seeger B (2008) Cluster kernels: resource-aware kernel density estimators over streaming data. IEEE Trans Knowl Data Eng 20(7):880–893
9. Parthasarathy S, Subramonian R (2001) An interactive resource-aware framework for distributed data mining. IEEE technical committee on distributed processing letters, pp 24–32
10. Shah R, Krishnaswamy S, Gaber MM (2005) Resource-aware very fast K-means for ubiquitous data stream mining. In: 2nd international workshop on knowledge discovery in data streams, pp 40–50
11. Tan PN, Kumar V, Srivastava J (2000) Indirect association: mining higher order dependencies in data. In: 4th European conference on principles of data mining and knowledge discovery, pp 632–637
12. Tan PN, Kumar V (2001) Mining indirect associations in web data. In: 3rd international workshop on mining web log data across all customers touch points, pp 145–166
13. Teng WG, Chen MS, Yu PS (2004) Resource-aware mining with variable granularities in data streams. In: 5th SIAM conference on data mining, pp 22–24

Chapter 3
Graph-Based Batch Mode Active Learning

Cheong Hee Park

Abstract Active learning aims to overcome the shortage of labeled data by obtaining class labels for some selected unlabeled data from experts. However, the selection process for the most informative unlabeled data samples can be demanding when the search is performed over a large set of unlabeled data. In this paper, we propose a method for batch mode active learning in graph-based semi-supervised learning. By acquiring class label information about several unlabeled data samples at a time, the proposed method reduces time complexity while preserving the beneficial effects of active learning. Experimental results demonstrate the improved performance of the proposed method.

Keywords Active learning · Batch mode active learning · Label propagation · Semi-supervised learning

3.1 Introduction

Semi-supervised classification utilizes useful information from unlabeled data to improve classification performance [1, 2]. Active learning is another possible approach when the number of labeled data samples is very small [3–6]. By requiring the exact class label of an unlabeled data sample from an expert, active learning enlarges the size of a labeled data set on which a new classifier is trained. This process is repeated by moving selected unlabeled data samples to a labeled data set one at a time. Because both active learning and semi-supervised learning aim to overcome the shortage of labeled data samples in different approaches, combining the two methods can be very powerful, as they can complement each other.

C. H. Park (✉)
Department of Computer Science and Engineering, Chungnam National University, 220, Gung-dong, Yuseong-gu, Daejeon 305-764, Korea
e-mail: cheonghee@cnu.ac.kr

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_3,

A method combining graph-based semi-supervised learning with active learning has been proposed in [7]. It selects an unlabeled data sample for which the knowledge of class label will cause the greatest reduction in the estimated error for the label prediction of other unlabeled data samples [7]. However, in every selection step, the expected prediction errors of unlabeled data samples should be computed for all possible labeling conditions, causing computational complexities to become burdensome. Additionally, the method only considers two-class problems and is limited to the framework of a first-order Gaussian Markov random [8].

In this paper, we extend the method in [7] to a method for batch mode active learning which is applicable to multi-class problems, using label propagation based on a second order Gaussian Markov random field. Instead of querying class labels of unlabeled data samples one at a time, class labels from several data samples are obtained at once, thus reducing the time complexity while maintaining the beneficial outcomes of active learning.

In Sect. 3.2, a batch mode active learning method for graph-based semi-supervised learning is presented, and in Sect. 3.3, experimental results demonstrate the performance of the proposed method. Discussion follows in Sect. 3.4.

3.2 Batch Mode Active Learning Based on Linear Neighborhood Propagation Method

We extend the active learning method in [7] by applying it to the Linear neighborhood propagation method (LNP) proposed in [8]. Suppose that a data set

$$X = \{x_1, x_2, \ldots, x_l, x_{l+1}, \ldots, x_{l+u}\}$$

is given. Let $x_i (1 \leq i \leq l)$ denote a labeled data sample having a class label y_i, and let $z_i \equiv x_{l+i} (1 \leq i \leq u)$ denote an unlabeled data sample whose class label y_{l+i} is unknown. For now, we assume two-class problems, in other words, a class label y_i equals to 0 or 1. Let $W = \left\{w_{ij}\right\}_{\{1 \leq i,j \leq l+u\}}$ be a similarity matrix for which w_{ij} represents similarities between data samples x_i and x_j. LNP defines the increment between a data sample and its neighboring data samples as $d_i = \sum_{j \in N_i} \left(w_{ij} y_j - y_i\right)^2$. This method induces the energy function under a second-order Gaussian Markov random field framework such as

$$E(y) = \frac{1}{2} \sum_i \left(y_i - \sum_{j \in N_i} w_{ij} y_j \right)^2 . \tag{3.1}$$

The solution that minimizes Eq. (3.1) is computed by the following:

$$f_u = -Q_{uu}^{-1} Q_{ul} f_l, \tag{3.2}$$

where $f_l = (y_1, \ldots, y_l)^T$, $f_u = (y_{l+1}, \ldots, y_{l+u})^T$ and $Q = (I - W)^T(I - W)$.

Suppose that the class label of unlabeled datum z_k becomes available as y_k. Then, the size of the labeled data set increases by one, and the size of the unlabeled data set decreases by one. The data set is reordered by exchanging z_1 and z_k, producing $\{x_1, \ldots, x_l, z_k, z_2, \ldots, z_{k-1}, z_1, z_{k+1}, \ldots, z_u\}$. Let W^* and Q^* be matrices obtained by permuting the rows and columns corresponding to the unlabeled data sample z_1 and the first unlabeled sample z_k in W and Q. If P denotes the permutation matrix exchanging the $(l + 1)$-th row and the $(l + k)$-th row of an identity matrix of size $(l + u)$ x $(l + u)$, then we obtain

$$\begin{aligned} Q^* = PQP = P(I - W)^T(I - W)P = (I - PWP)^T(I - PWP) \\ = (I - W^*)T(I - W^*) \end{aligned}$$

Let Q^* be partitioned as

$$Q^* = \begin{bmatrix} Q^*_{ll} & Q^*_{lu} \\ Q^*_{ul} & Q^*_{uu} \end{bmatrix}, \quad Q^*_{ul} = \begin{bmatrix} d \\ E \end{bmatrix}, \quad Q^*_{uu} = \begin{bmatrix} a & b \\ b^T & R_{u-1,u-1} \end{bmatrix} \tag{3.3}$$

> For notational convenience, small letters represent scalars and row or column vectors and capital letters are reserved for matrices. Then, $\left(Q^*_{uu}\right)^{-1}$ is obtained by permuting the first and the k-th rows and columns of Q^{-1}_{uu}, and we write $\left(Q^*_{uu}\right)^{-1}$ as

$$\left(Q^*_{uu}\right)^{-1} = \begin{bmatrix} x & v \\ v^T & Z \end{bmatrix}. \tag{3.4}$$

Now, by Eq. (3.2), when z_k is moved to the labeled data set, the class labels of other unlabeled data samples are predicted as follows:

$$f_u^{-(z_k, y_k)} = -R^{-1}_{u-1,u-1} \begin{bmatrix} E & b^T \end{bmatrix} \begin{bmatrix} f_l \\ y_k \end{bmatrix} \tag{3.5}$$

The inversion of a matrix in block form can be computed by the following equation:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

where A and D are invertible [9]. In particular, when [A B] is a row vector and the matrix is symmetric, the inverse is given by

$$\begin{bmatrix} a & b \\ b^T & D \end{bmatrix}^{-1} = \begin{bmatrix} \tau & -\tau bD^{-1} \\ -\tau D^{-1}b^T & D^{-1} + \tau D^{-1}b^T bD^{-1} \end{bmatrix} \tag{3.6}$$

where $\tau = \frac{1}{a - bD^{-1}b^T}$ and $\left(D - \frac{1}{a}b^T b\right)^{-1} = D^{-1} + \tau D^{-1}b^T bD^{-1}$.

By applying Eq. (3.6) to Q^*_{uu} contained in Eq. (3.3) and by comparing its components with those of $(Q^*_{uu})^{-1}$ in Eq. (3.4), we have

$$x = \tau, where \tau = \frac{1}{a - bR^{-1}_{u-1,u-1}b^T}$$

$$v^T = -\tau R^{-1}_{u-1,u-1}b^T,$$

$$Z = R^{-1}_{u-1,u-1} + \tau R^{-1}_{u-1,u-1}b^T bR^{-1}_{u-1,u-1} = R^{-1}_{u-1,u-1} + \frac{1}{x}v^T v.$$

Hence, after deleting one row and one column from Q^*_{uu}, the inverse of $R_{u-1,u-1}$ can be expressed in terms of submatrices of $\left(Q^*_{uu}\right)^{-1}$:

$$R^{-1}_{u-1,u-1} = Z - \frac{1}{x}v^T v \tag{3.7}$$

By substituting $R^{-1}_{u-1,u-1}$ in (3.5) by its representation in Eq. (3.7), we complete the formula for the label prediction of unlabeled data samples:

$$f_u^{-(z_k,y_k)} = (P_k f_u)(2{:}u) + \frac{v^T}{x}(y_k - f_u(k))$$

Here, P_k is the permutation matrix exchanging the first row and the k-th row of an identity matrix of size $u \times u$ and $(P_k f_u)(2{:}u)$ denotes a submatrix composed of the second row through the u-th row of $P_k f_u$.

For a multi-class problem, the class label of a data sample belonging to class $i \in \{1, \cdots, c\}$ is represented as a row vector whose i-th component is one and all others are zero. Then f_u in Eq. (3.2) becomes a $u \times c$ matrix. For the k-th unlabeled data sample z_k, we have

$$H_{f_u}(z_k, i) = \frac{f_u(k,i) - \min\limits_{1 \le j \le c} f_u(k,j)}{\sum_{s=1}^{c}\{f_u(k,s) - \min\limits_{1 \le j \le c} f_u(k,j)\}}$$

This equation can be interpreted as representing the confidence with which z_k will belong to class i; therefore, the sample is predicted to belong to the class for which the sample has the highest confidence. When the class label of z_k is y_k, the estimated prediction error of $f_u^{-(z_k,y_k)}$ can be defined as follows:

$$R\left(f^{-(z_k,y_k)}\right) = \sum_{z_j \in \text{unlabeled}} \left(1 - \max_{1 \le i \le c} H_H f_u{}^{-(z_k,y_k)}(z_j,i)\right).$$

Additionally, the expected estimated error $R(f^{-z_k})$ over all the possible labeling of z_k is defined by the following equation:

$$\mathrm{R}(\mathrm{f}^{-\mathrm{z_k}}) = \sum_{j=1}^{c} H_{f_u}(\mathrm{z_k}, j) R\left(f_u^H f_u(\mathrm{z_k}, j) R\left(f_u^{-(\mathrm{z_k}, j)}\right)\right)$$

The easiest way to carry out batch mode active learning is by selecting several samples having the smallest values of $\mathrm{R}(\mathrm{f}^{-\mathrm{z_k}})$; however, this method carries the risk of selecting samples belonging to a single class. A balanced selection across different classes can prevent information overlap and facilitate the independence of selected samples. To achieve a balanced selection from each class, we divide the unlabeled data set into c sets $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_c$ such as

$$\mathrm{A_i} = \{\mathrm{z_k} \in \text{unlabeled} \,|\, \mathrm{i} = \arg\max_{\mathrm{j}} \mathrm{f_u}(\mathrm{k}, \mathrm{j}).$$

We then select the data sample having the smallest expected estimated error $\mathrm{R}(\mathrm{f}^{-\mathrm{z_k}})$ from each $\mathrm{A_i}$ and acquire the class labels of c selected data samples in a single step.

3.3 Experimental Results

To test our proposed method, we conducted experiments using real data sets. Four data sets, image segment, optical digit, satellite, and mfeature, were downloaded from the UCI machine learning repository. The detailed description is given in Table 3.1. For each data set, the labeled data set was composed by randomly selecting one data sample from each class, and the remaining was assumed to be unlabeled data.

We compared batch mode active learning with active learning extended from the method in [4] as well as with active learning by random sampling. Whenever the size of the labeled set was increased by active learning, we measured the prediction accuracy of remaining unlabeled data samples. Prediction accuracies were averaged after repeating ten random trials in the construction of initial labeled and unlabeled data sets. The average prediction accuracies over the iterations of the selection process are displayed in Fig. 3.1. The number of added labeled data samples are marked on the x-axis, and prediction accuracies are shown on the y-axis. The original active learning algorithm selects unlabeled data samples one by one, but our batch mode active learning algorithm selects several unlabeled data samples at once. For example, in the optical digit set, to select 100

Table 3.1 Description of data sets

Data set	Samples	Classes	Features
Image segment	2,310	7	19
Optical digit	5,620	10	64
Satellite	6,435	6	36
mfeature	2,000	10	256

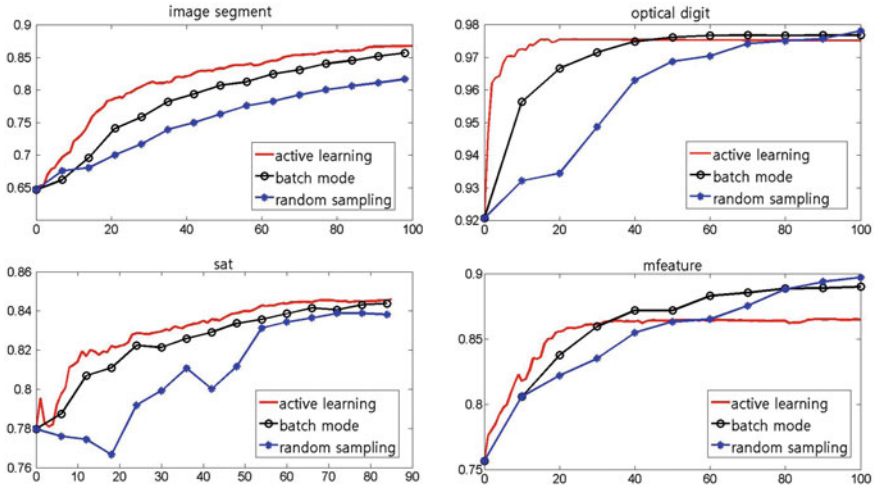


Fig. 3.1 The average prediction accuracies over the iterations of the selection process. The number of labeled data samples added is marked on the x-axis, and prediction accuracies are shown on the y-axis

unlabeled data samples, the original active learning should repeat the selection process 100 times; in contrast, in batch mode active learning, 100 unlabeled data samples are selected in only ten iterations of the selection process. Hence, implementing batch mode active learning can reduce computing time over active learning by a large factor. For that matter, as shown in Fig. 3.1, prediction accuracies in batch mode active learning approach those of the original active learning after sufficient iteration.

3.4 Discussion

In the proposed batch mode active learning method, c data samples are chosen in each selection stage. In contrast, the original active learning method selects one data sample in one time, therefore repeating the steps for one sample selection. Hence, the proposed batch mode algorithm can reduce the computational time complexity of selection process by a factor of c. In particular, when the scale of the unlabeled data set is large, the effect of time complexity reduction becomes great.

References

1. Zhu X (2006) Semi-supervised learning literature survey. Technical report, Computer sciences, University of Wisconsin-Madison
2. Chapelle O, Scholkopf B, Zien A (2006) Semi-supervised learning. MIT press, Cambridge
3. Li M, Sethi I (2006) Confidence-based active learning. IEEE Trans Pattern Anal Mach Intell 28(8):1251–1261
4. Freund Y, Seung HS, Shamir E, Tishby N (1997) Selective sampling using the query by committee algorithm. Mach Learn 28(2–3):133–168
5. Cohn DA, Ghahramani Z, Jordan MI (1996) Active learning with statistical models. J Artif Intell Res 4:129–145
6. Settles B (2009) Active learning literature survey. Computer sciences technical report 1648. University of Wisconsin-Madison
7. Zhu X, Ghahramani Z, Lafferty J (2003) Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of ICML, pp 912–919
8. Wang J, Wang F, Zhang C, Shen H, Quan L (2009) Linear neighborhood propagation and its application. IEEE Trans Pattern Anal Mach Intell 31(9):1600–1615
9. Simon D (2006) Optimal state estimation. Wiley

Chapter 4
One Pass Outlier Detection for Streaming Categorical Data

Swee Chuan Tan, Si Hao Yip and Ashfaqur Rahman

Abstract Attribute Value Frequency (AVF) is a simple yet fast and effective method for detecting outliers in categorical nominal data. Previous work has shown that AVF requires lesser processing time while maintains very good outlier detection accuracy when compared with other existing techniques. However, AVF works on static data only; this means that AVF cannot be used in data stream applications such as sensor data monitoring. In this paper, we introduce a modified version of AVF known as One Pass AVF to deal with streaming categorical data. We compare this new algorithm with AVF based on outlier detection accuracy. We also apply One Pass AVF for detecting unreliable data points (i.e., outliers) in a marine sensor data monitoring application. The proposed algorithm is experimentally shown to be as effective as AVF and yet capable of detecting outliers in streaming categorical data.

Keywords Data stream · Outlier · Categorical data · Attribute value frequency · One pass

4.1 Introduction

Outlier detection is the process of detecting instances with unusual behavior that occurs in a system. Effective detection of outliers can lead to the discovery of valuable information in the data. Over the years, mining for outliers has received

S. C. Tan (✉) · S. H. Yip
School of Business, SIM University, 535A Clementi Road, Singapore, Singapore
e-mail: jamestansc@unisim.edu.sg

S. H. Yip
e-mail: shyip002@unisim.edu.sg

A. Rahman
Intelligent Sensing and Systems Laboratory, CSIRO, Hobart, Australia
e-mail: ashfaqur.rahman@csiro.au

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_4,

significant attention due to its wide applicability in areas such as detecting fraudulent usage of credit cards, unauthorized access in computer networks, weather prediction and environmental monitoring.

A number of existing methods are designed for detecting outliers in continuous data. Most of these methods use distances between data points to detect outliers. In the case of data with categorical attributes, attempts are often made to map categorical features to numerical values. Such mappings impose arbitrary ordering of categorical values and may cause unreliable result.

Another issue is related to the big data phenomenon. Many systems today are able to generate and capture real-time data continuously. Some examples include real-time data acquisition systems, condition monitoring systems, and sales transaction systems. It is a challenging task to effectively detect outliers occurring in data streams. Traditional outlier detection approaches are no longer feasible as they only deal with statics data sets and require multiple scans of data to produce effective results. In data streams setting, outlier detection algorithms (e.g., [8]) need to process each data item within a strict time constraint and can only afford to analyze the entire data set with a single scan of data.

In this paper, we introduce a modified version of Attribute Value Frequency (denoted as AVF, which is proposed by Koufakou et. al [6]) for outlier detection. This version is known as One Pass Attribute Value Frequency (One Pass AVF) for outlier detection in streaming categorical data. Note that AVF computes the frequency of each attribute-value pair in the *entire* data set. In contrast, One Pass AVF computes the *cumulative probability* of each possible attribute-value pair that has been identified at the time point of processing a data stream. As a result, One Pass AVF is capable of detecting outliers in just a single scan of the data, and this allows it to process massive streaming data. The focus of this paper is twofold (i) to compare One Pass AVF with the original AVF based on detection accuracy, and (ii) to apply One Pass AVF in a real world marine sensor data monitoring application.

4.2 Related Work

Most existing outlier detection methods are designed to work on numerical data. For example, distance-based and density-based techniques cannot handle data sets containing categorical attributes because the notions of ‘distance’ or ‘density’ are not well defined. Yet categorical data is commonly found in many real-world databases. In the following, we will review some of the existing methods for mining categorical outliers.

Frequent Pattern Outlier Factor (*FPOF*): This method [4] uses association rule mining technique (e.g., [1]) to find frequently occurring itemsets in data. It then assigns an outlier score to each data point based on the number of frequent itemsets associated with the data point. As real-world data sets are usually large, this approach requires longer processing time to find frequent itemsets. In addition,

it requires many attempts to locate an appropriate support threshold for identifying frequent itemsets.

Hypergraph based Outlier Test: The Hypergraph based Outlier Test (HOT) was proposed [11] to deal with large data set with missing values and mixed-type attributes. HOT uses connectivity to process data with missing values and its detection results are easy to interpret. However, HOT cannot be extended to deal with streaming data.

Attribute Value Frequency: Most traditional outlier detection approach like the Greedy algorithm [5] requires multiple scans of the data to produce an effective result. Indeed, these methods tend to slow down when the data set becomes large. To address this problem, Koufakou et. al. [6] proposes the use of Attribute Frequency Value method to detect outliers that have few occurrences and irregular attribute values. The outlier score, AVF, is obtained by computing the relative frequency of occurrences of attribute-value. It is formulated as $\text{AVFScore}(x_i) = 1/m \cdot \sum f(x_{ij})$; where m is number of categorical variables, the summation runs from $j = 1, 2, \ldots, m$, and $f(x_{ij})$ is the relative frequency of the jth attribute value of instance x_i appearing in the data set. Data points that are few and different [10] will contain lower AVF scores, and tend to have higher probability of being outliers.

As AVF algorithm requires less scans of data to identify outliers, it is significantly faster than many existing methods. Unlike FPOF, AVF is easy to implement since it does not generate frequent itemsets; it is also easier to use since it does not require users to set the minimum support threshold.

One major drawback of AVF is that it works on static data sets only. The algorithm needs to load the entire data set into the computer memory and then scan through all the records, before it can start detecting outliers. Fortunately, it is quite easy to extend AVF to deal with streaming data. In the following section, we will present such an extension of AVF.

4.3 Proposed Method

We name the proposed extension of AVF as One Pass Attribute Value Frequency (One Pass AVF) method for detecting streaming categorical outliers. Note that One Pass AVF uses cumulative probability of each attribute-value pair in *instances seen so far in the data stream*, whereas AVF computes the relative frequency of each attribute-value pair in the *entire* data set. The use of cumulative probability allows One Pass AVF to perform one-pass outlier detection in constant time and with fixed amount of memory space when processing each instance. In data stream environment, instances flow in continuously and there is no room to store the whole data set, and the algorithm cannot perform multiple accesses of the data instances. As each instance streams in, One Pass AVF processes it and then disposes the instance right away, thereby preventing the memory from running out of space. One Pass AVF score is formulated as shown: $\text{OPAVFScore}(x_i) = 1/m \cdot \sum p(x_{ij})$, where m is number of categorical variables, and $p(x_{ij})$ is cumulative

probability of the jth attribute value of instance x_i in the data stream. The summation runs from $j = 1, 2, \ldots, m$.

Algorithm 1 shows the proposed One Pass AVF algorithm. As each streaming data point x is being read in, the cumulative probability of each attribute-value pair found in x is computed. Then, the OPAVFScore of x is being computed and reported. Finally, x is being discarded to release computer memory before processing the next streaming instance.

Consider a toy data set with one nominal attribute A and four observations: {'p', 'p', 'q', 'p'}. In the beginning, Algorithm 1 reads in the first streaming data point 'p', with frequency of occurrence equals 1 (i.e., f(A = 'p') = 1), and the cumulative probability of 'p' occurring equals 1 (i.e., p(A = 'p') = 1/1). For the second streaming point 'p', the frequency of occurrence is 2 and its cumulative probability of occurrence is 1 (since p(A = 'p') = 2/2 = 1). For the third streaming point, the frequency of occurrence for 'q' is 1 and its cumulative probability of occurrence is 1/3. The fourth and last streaming point 'p' has frequency of occurrence equals 3 and the cumulative probability of occurrence equals 3/4. At the end of the process, a list of OPAVFScores containing {1, 1, 1/3, 3/4} is produced. With respect to the stream's sequence, it correctly assigns the third observation 'q' with the lowest OPAVFScore of 1/3, signifying it to be an outlier.

Algorithm 1: The proposed One Pass AVF Algorithm. f(A = c) is the number of times that the event A = c has occurred so far. p(A = c) is the cumulative probability of event A = c.

Algorithm 1: The proposed One Pass AVF Algorithm. f(A=c) is the number of times that the event A=c has occurred so far. p(A=c) is the cumulative probability of event A=c.

```
Algorithm OnePassAVF
begin
  count <- 0
  while data stream continues do
    Read in the next streaming data point x
    count++
    for each attribute A in x do
      for each category c in A do
        if A = c then
          p(A = c) <- f(A = c)/count
        end if
      end for
    end for
    Report OPAVFScore (x) as the anomaly score for x
    Discard x from computer memory
  end while
end.
```

4.4 Experimental Setup

The data sets used in this project are summarized in Table 4.1. There are two experiments. The first experiment involves comparing One Pass AVF and AVF using the four data sets taken from UCI Machine Learning Repository [2], namely Post-Operative, Lymphography, Breast Cancer Wisconsin and Page Block. The first three data sets contain categorical attributes, whereas the continuous attributes in the Page Block data set are discretised as categorical type. These datasets are good for benchmarking because AVF is known to perform very well on these data sets [6].

For each method being tested, the evaluation of its detection accuracy starts after the anomaly scores of all instances in the data set have been collected. First, the first 10 % of the streaming instances are excluded because the One Pass AVF model is still not fully developed during the initial stage. The remaining 90 % of the streaming data are then ranked based on their anomaly scores. From this ranking and the ground truth, we compute the AUC (Area Under receiver operating characteristic Curve) [3] to measure the performance of the method. A higher AUC score means better detection accuracy.

The second experiment involves evaluating the effectiveness of One Pass AVF in a real-world sensor data monitoring application [7, 9]. In this experiment, we use One Pass AVF to rank the quality of streaming sensor data obtained from conductivity sensors located in Sullivan Cove, Hobart, Australia. The conductivity sensors are among other sensors being installed as part of the Tasmanian Marine Analysis Network Project [7, 9] to monitor the condition of the waterways in south-east Tasmania. Over the years, growing activities in shipping, urban development, aquaculture and other uses of waterways in the region are imposing pressure to the health of waterways. As a result, it is important to assess the condition of the waterways to understand the impact of these activities.

To monitor the condition of the waterway accurately, it is necessary to ensure that the sensor data is of high quality. Factors that could degrade sensor performance include algae accumulation on the sensor surface, seasonal variation, and drift in the sensor's electrical output. The quality of the sensor data has been assessed and labeled by two domain experts, and there are four categories: (1) Good, (2) Probably good, (3) Probably bad and (4) Bad. These labels are then used as ground truth to assess the performance of One Pass AVF.

Tab 4.1 Summary of data sets used in the experiments

Data set	No. of instances	No. of attributes	Classes considered as outlier	% of outliers
Post-operative	90	10	Classes 1 and 2	29
Lymphography	148	19	Classes "Fibrosis" and "Normal"	4
Wisconsin breast cancer	483	10	Class 4	8
Page block	5,473	11	Classes 2, 3, 4 and 5	10.2
Shall_Cond	19,375	12	Quality Labels 3 and 4	1.2

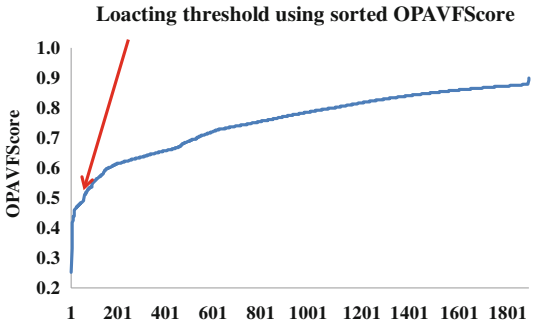


Fig. 4.1 The threshold is determined based on the middle of the 'bend'. This threshold is used for distinguishing outliers from normal instances using One Pass AVF

The data set (denoted as Shall_Cond) contains observations from the "Shallow Conductivity" sensor placed at one metre below the lowest possible astronomical low tide at the location of deployment. The sampling interval is 5 min and data was collected from 19 Feb 2008 to 08 Nov 2010. It is interesting to evaluate the performance of One Pass AVF on the Shall_Cond data set for two reasons. Firstly, the Shall_Cond data set is known to contain a very low frequency of unreliable data (i.e., observations that are 'Probably bad' or 'Bad'). Secondly, this data set is collected from a real data stream environment that requires one pass data processing.

To use One Pass AVF in a real-world sensor data monitoring application, continuous attributes in this data set are discretised (into ten equal-width bins) and the first 10 % of the streaming instances are used to construct the initial model. In addition, this initial 10 % of the instances are also used to define a threshold that will be used to distinguish outliers from normal instances. Instances with outlier scores greater than the threshold are classified as normal; otherwise they are classified as outliers.

The threshold is determined by locating the 'bend' of the sorted OPAVFScore profile as shown in Fig. 4.1. Notice that before this 'bend', the OPAVFScores are relatively low but increasing sharply, indicating that the associated instances contain less frequent attribute-value pairs (i.e., lower probability of occurrences). After this 'bend', the OPAVFScores are relatively high and the rate of increase in the values becomes a lot slower, suggesting that the instances are less distinctive and are more likely to be normal instances with higher frequency of occurrences. For the Shall_Cond data set, the threshold is set at 0.53, which is in the middle of the bend identified in Fig. 4.1.

4.5 Results and Discussions

Comparing One Pass AVF and AVF: Table 4.2 shows the AUC scores achieved by AVF and One Pass AVF. A model with high AUC is expected to identify outliers more effectively. From the table, it is clear that the AUC scores of One Pass AVF

Tab 4.2 AUC scores of AVF and One Pass AVF when applied to four data sets

Area under ROC curve		
Data set	AVF	One Pass AVF
Post-operative	0.757	**0.789**
Lymphography	**0.990**	0.956
Wisconsin breast cancer Wisconsin	0.987	0.987
Page block	0.679	**0.681**

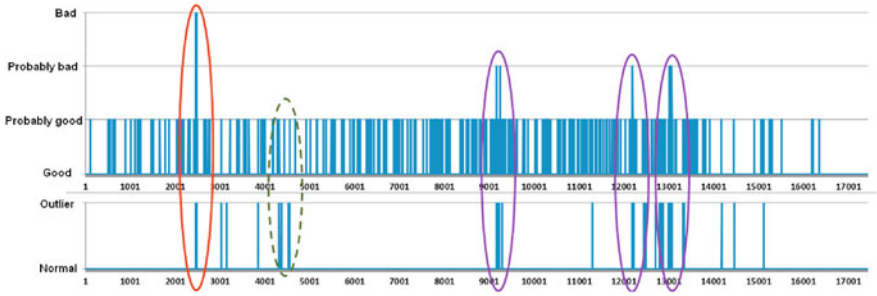


Fig. 4.2 One Pass AVF for sensor data monitoring. The x-axis shows the time points. The *top plot* shows the actual data quality at various time points (i.e., the ground truths). The *bottom plot* shows the notifications of outlier occurrences detected by One Pass AVF

are very close to AVF. In fact, One Pass AVF outperforms AVF on the Post-Operative and Page Block data sets, and it has the same score as AVF on the Wisconsin Breast Cancer data set. This shows that One Pass AVF, despite using less information, has performance that is comparable to AVF.

Applying One Pass AVF in Marine Sensor Data Monitoring: Figure 4.2 displays two plots used to evaluate the effectiveness of One Pass AVF for detecting unreliable sensor data. The top plot shows the occurrences of streaming instances (after the first 10 % of streaming data) along with the ordinal quality label of the sensor data along the vertical axis ("Good", "Probably good", "Probably bad", and "Bad"). The bottom plot shows the time points where One Pass AVF notifies the occurrences of outliers along with two categories ("Normal", and "Outlier") along its vertical axis.

The top plot shows that it may not be practical and necessary to raise alerts for instances with quality label "Probably good" because such instances occur almost throughout the entire stream. Instead, focus should be on the "Probably bad" and "Bad" instances, where actions should be taken when such outliers occur.

At around the time point of 2500, there are a total of 11 outliers detected. The occurrences of these "Bad" instances are all detected by One Pass AVF. Subsequently, the occurrences of "Probably bad" instances at around the time points of 9,000, 12,000 and 13,000, are also being identified by One Pass AVF. Note that there are a few false alarms as well. One example is at around 4500th time point where there are indeed no poor quality label issues.

4.6 Concluding Remarks

In this paper, we have compared the proposed One Pass AVF method against AVF using both categorical and continuous data sets. Experimental results have shown that the outlier detection accuracy of One Pass AVF is comparable to that of AVF. We also applied One Pass AVF in an automated sensor data quality assessment and evaluate its effectiveness for identifying outliers. Experimental results have shown that the algorithm is able to detect all occurrences of data quality issues (Quality labels "Probably bad" and "Bad") in the sensor data. We also noted that there are a few cases of false alarms during the detection process. One possibility for future work is to identify the root cause of such occurrences so that the model may be improved to minimize such cases. Finally, readers interested in this work (e.g., data sets) can contact the first author. Other details of this work will also be available online.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In Poceedings of the International Conference on Very Large Data Bases VLDB, pp. 487–499, 1994
2. Asuncion A, Newman DJ (2007) UCI machine learning repository. University of California, Irvine, CA. http://archive.ics.uci.edu/ml
3. Hand DJ, Till RJ (2001) A simple generalisation of the area under the roc curve for multiple class classification problems. Mach Learn 45:171–186
4. He Z, Xu X, Huang JZ, Deng S (2005) FP-Outlier: frequent pattern based outlier detection. Comput Sci Inf Syst 2(1):103–118
5. He Z, Xu X, Deng S (2005) An optimization model for outlier detection in categorical data. In: Proceedings of the 2005 international conference on advances in intelligent computing—volume part I. Springer, Berlin, Heidelberg, pp 400–409
6. Koufakou A, Ortiz E, Georgiopoulos M, Anagnostopoulos G, Reynolds K (2007) A scalable and efficient outlier detection strategy for categorical data. In: IEEE international conference on tools with artificial intelligence ICTAI, pp 210–217
7. Rahman A, Smith D, Timms G (2011) Multiple classifier system for automated quality assessment of marine sensor data. In: Proceedings of the twenty-second international joint conference on artificial intelligence, pp 1511–1516
8. Tan SC, Ting KM, Liu FT (2013) Fast anomaly detection for streaming data. In: Proceedings of IEEE intelligent sensors, sensor networks and information processing (ISSNIP), pp 362–367
9. Timms GP, McCulloch JW, McCarthy P, Howell B, de Souza PA, Dunbabin MD, Hartmann K (2009) The Tasmanian marine analysis network (TasMAN). In: Proceedings of IEEE oceans, vol ½. Bremen, Germany, pp 43–48
10. Ting KM, Zhou GT, Liu FT, Tan SC (2013) Mass estimation. Mach Learn 90(1):127–160
11. Wei L, Qian W, Zhou A, Jin W, Yu J (2003) HOT: hypergraph-based outlier test for categorical data. In: Proceedings of the 7th Pacific-Asia conference on advances in knowledge discovery and data mining, pp 399–410

Chapter 5
Measuring of QoE for Cloud Applications

Yu-Hui Tao, Yu-Lung Wu, Chi-Jui Chang and Chi-Wen Chang

Abstract According to NIST [3], cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources, such as networks, servers, storage, applications, and services, that can be rapidly provisioned and released with minimal management effort or service provider interaction. Three service models are included in this cloud model, including cloud Software as a Service (SaaS), cloud Platform as a Service (PaaS), and cloud Infrastructure as a Service (IaaS). Furthermore, the proposed applications of cloud computing may be related to consumer electronics, such as virtualization of consumer storage, and Cloud TV platforms that provide access to a number of Web applications, such as social networking, user generated video games, etc. [4].

Keywords Cloud computing · Quality of experience · Quality of service · Decision tree · Experiential marketing

Quality of service (QoS) plays an important role in cloud-based services. In other words, to achieve end-to-end QoS solutions, IP network providers will need to agree on a common set of IP packet transfer performance parameters and QoS objectives. An end-to-end IP QoS solution enabling successful IP/PSTN convergence will likely be realized in three steps [6]: achieving network provider agreement, deploying network mechanisms, and embedding the QoS objectives in signaling protocols to enable on-demand creation of QoS-assured IP flows.

Y.-H. Tao
Department of Information Management, National University of Kaohsiung, Kaohsiung, Taiwan R.O.C

Y.-L. Wu
Department of Information Management, I-Shou University, Kaohsiung, Taiwan R.O.C

C.-J. Chang · C.-W. Chang (✉)
Department of Information Engineering, I-Shou University, Kaohsiung, Taiwan R.O.C
e-mail: hellocavin1021@gmail.com

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_5,

While QoS represents an objective network-oriented measure of efficiency when providing a service, typically expressed with delay, jitter, lost information, and throughput parameters, quality of experience (QoE) is a subjective and network independent measure of service efficiency as perceived by the end-user when consuming the service as well as a measure of ability of the system to achieve the end-user's expectations [7]. Both service delivery and customer satisfaction are critical in modern commercial management platform. Particularly, [2] indicated that QoE factors have to be considered as key input for a successful business operation between a customer and a company. Meanwhile, [1] pointed out that QoE as perceived by users has the potential to become the guiding paradigm for quality in the cloud.

However, the growing presence of cloud-based services creates new problems for both users and providers, resulting in a number of challenges that need to be addressed in order to ensure successful adoption of this new paradigm [1]. For instance, a cloud service solution in remote data centers is an emerging means of providing user in an easy-to-maintain way. The fact that such solutions rely on the presence of connectivity between end users and remote data centers poses a challenging question of a quality of experience for end users. What is the quality of experience of the user when running a particular application in cloud services? The challenge is to understand whether the end users perceived between the client and the server has enough characteristics to satisfy the end users.

In this article, we will discuss technical challenges to understand user's perceived factors from the quality of experience as perceived by users, as well as how those factors impact QoE. The traditional approach in the literature using the QoS perspective or IP parameters for the parameters of QoE is inadequate. For example, in certain QoS situations when the quality parameters have shown inferior quality, the user QoE may be satisfactory due to factors other than QoS scope. Therefore, a new and comprehensive perspective of QoE for Cloud services is desirable.

To address this question, we plan to first use factors in the dimensions of Experiential Marketing by [5], including sense, feel, think, act, and relate to develop a scale for measuring the QoE of cloud-base service. Accordingly, we propose a method that exploits statistical classification to infer the key factors of the measurable parameters of the QoE for a given cloud service application, such as data-, voice- or video-related services that may be for mobile web, web, or installed agents. A review of related literature and industrial practices will be conducted to derive a two-dimension table of data type (data, voice, video) versus of applications (mobile, web, and installed agent). Statistical package of SPSS will be used for this stage of analyses.

We then correlate such information with the key factor to obtain the user's QoE. We evaluate how machine-learning techniques, such as decision tree C4.5, can robustly detect the QoE of the cloud services that can satisfy the end users, with promising results. Weka (http://www.cs.waikato.ac.nz/ml/weka) will be used for this stage of analyses. The target sample will be the customers of a large telecommunication corporation in Taiwan and the sample will be large enough for

meeting the needs of data mining techniques. Together with a QoE-based classification model of cloud service applications using the existing customers of a large telecommunication corporation, these can drive the research agenda on QoE management for practical cloud applications.

To the best of our knowledge, this is the first attempt of using user centered statistical techniques to measure the Cloud services for QoE detection, which will contribute certain new knowledge to the research areas of QoE and Cloud-based Services.

References

1. Hobfeld T, Schatz R, Varela M, Timmerer C (2012) Challenges of QoE management for cloud applications. IEEE Commun Mag 50(4):28–36
2. Laghari KUR, Crespi N, Molina B, Palau CE (2011) QoE aware service delivery in distributed environment, advanced information networking and applications (WAINA). IEEE Workshops of International Conference on Digital Object Identifier, pp. 837–842
3. Mell P, Grance T (2011) The NIST definition of cloud computing: recommendations of the national institute of standards and technology. NIST Spec Publ 145(6):7
4. Sánchez R, Almenares F, Arias P, Díaz-Sánchez D, Marín A (2012) enhancing privacy and dynamic federation in IdM for consumer cloud computing. IEEE Trans Consum Electron 58(1):95–103S
5. Schmitt, Bernd H (1999) Experiential marketing: how to get customers to sense, feel, think, act, and relate to your company and brand*s*, free press, 1st edn. Simon & Schuster, New york
6. Seitz N (2003) ITU-T QoS standards for IP-based networks. IEEE Commun Mag 41(6):82–89
7. Sterle J, Volk M, Sedlar U, Bester J, Kos A (2011) Application—based NGN QoE controller. IEEE Commun Mag 49(1):92–100

Chapter 6
Mining Weighted Partial Periodic Patterns

Kung-Jiuan Yang, Tzung-Pei Hong, Yuh-Min Chen and Guo-Cheng Lan

Abstract In the data mining area, partial periodic pattern mining has become an important issue in many business applications. Although weighted sequential pattern mining algorithms have been widely discussed, until now, there is no further discussion in the field of weighed partial periodic pattern mining. Thus, this work introduces a new research issue, named weighted partial periodic pattern mining, which considers the individual significances of events in an event sequence. In addition, a projection-based mining algorithm is presented to effectively handle the weighted partial periodic pattern mining problem. The experimental results show the proposed algorithm is efficient.

Keywords Data mining · Partial periodic pattern mining · Weighted partial periodic pattern mining

K.-J. Yang
Department of Information Management, Fortune Institute of Technology, Kaohsiung 831, Taiwan
e-mail: kjoyang@gmail.com

T.-P. Hong (✉)
Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung City 811, Taiwan
e-mail: tphong@nuk.edu.tw

Y.-M. Chen
Institute of Manufacturing Information and Systems, National Cheng Kung University, Tainan 701, Taiwan
e-mail: ymchen@mail.ncku.edu.tw

G.-C. Lan
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan
e-mail: rrfoheiay@gmail.com

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_6,

6.1 Introduction

Many studies have contributed to the efficient mining algorithms to improve efficiency and effectiveness in mining sequential patterns [4, 7, 8]. Several weight based sequential pattern mining algorithms were proposed as well [10–12]. The main task in weight based sequential pattern mining is how to keep the downward-closure property when applying the weight constraints [1]. Besides, periodic pattern mining is one of the important issues in sequential pattern mining to discover regularity in the time series or event sequences. Since not every activity in the real world has full regularity, it thus generates another kind of periodic pattern mining called partial periodic pattern mining, which is so-called "less restrictive" mining. The main concept of the partial periodic pattern mining was first introduced by Han et al. [5] to find partial periodicity could be associated with a subset of the time points of the periodic behavior. In other words, a partial periodic pattern is a mixture of the periodic events and non-periodic events.

Based on the above reasons, this study presents an efficient approach, projection-based weighted upper-bound partial periodic pattern mining algorithm (Abbreviated *PWA*), to adopt the maximum weight of upper-bound model proposed by Yun et al. [10] to prune the weighted infrequent partial periodic patterns and the efficient projection-based partial periodic pattern mining algorithm proposed by Yang et al. [9] to find the weighted partial periodic patterns with a specific period length in an event sequence. The *PWA* algorithm is designed to have the upper-bounds of maximum weight to the event tuples of each segment, and then to adopt projection-based algorithm to reduce candidates to quickly recognize weighted partial periodic patterns. Finally, the experimental results reveal that the good performance in discovering the meaningful partial periodic patterns.

6.2 Review of Related Works

There are many studies in mining weighted frequent pattern based on the pattern-growth approach have been developed to find the important patterns [2, 3]. For instance, Cai et al. defined a weighted support calculated by multiplying the support of a pattern with that pattern's average weight [3]. Besides, the k-support bound was designed to maintain the anti-monotone property in mining association rules with w items. Extending the weighted concept from finding weighted frequent patterns to weighted sequences, Yun et al. applied the maximum weight among items as maximum weight of each sequence to discover weighted sequential patterns from a sequence data [11, 12].

It is observed that the advantage of weight-based algorithms had been effectively applied in sequential pattern mining, but specialized in partial periodic pattern mining is never been discussed. Since partial periodic pattern mining is

popular in business applications, it motivates us to adopt a weight-based concept to find weighted partial periodic patterns. The supportive discussion will be made in the following section.

6.3 Problem Statement and Definitions

To explain the problem of weighted PPP (abbreviated WPPP) mining with a specific period length clearly, assume that there is an event sequence *"abcdabcbacad"* composed of twelve events in order of their time stamp. There are four different events in the sequence, denoted as event set E = {*a, b, c, d*}, and the weight values of the four events are given in the weight set W = {0.3, 0.8, 0.9, 0.5} respectively. The specific period length is set to 4 and the pre-defined minimum weighted support threshold *(min_wsup)* is set at 35 %. A set of terms related to WPPP mining with a specific period length is defined below.

An event sequence *S* is a time series of events and in the order of their occurrence time, denoted as $S = <e_1, e_2, \ldots, e_n>$, where e_j is the *j*-th event in *S*. Let $I = \{i_1, i_2, \ldots, i_k\}$ be a set of all events. For example, $S = < abcdabcbacad >$. The first and second event, *a* and *b*, respectively, appear in the first and second time stamps in *S*, respectively. According to the given a period length of *l*, an event sequence *S* can be divided into $(m = \frac{n}{l})$ mutually disjoint period segments. The event sequence *S* can then be denoted as $S = < ps_1, ps_2, \ldots, ps_m >$. Because the period length is set to 4, the event sequence *S* can be divided into three period segments $ps_1 = < abcd >$, $ps_2 = < abcb >$ and $ps_3 = < acad >$. The support value sup_{cp} of a candidate pattern *cp* with the period length of *l* is the count of the period segments including the *cp* over the total number of period segments. For example, the candidate pattern $< abcd >$ only appears in the first period segment, and then its support value is found as 1/(12/4) = 33.33 %.

In this study, an event tuple is composed of the event and its position identifier in the corresponding period segment ps_j of *S*. For example, the first period segment ps_1 can be encoded as $eps_1 = <(a,1), (b,2), (c,3), (d,4)>$. The support value sup_{ctp} of a candidate tuple pattern *ctp* is the count of encoded period segments including *ctp* over the number of encoded period segments in *S*.

The maximum period segment weight $mpsw_j$ of an encoded period segment eps_j is the maximum value of the weighted values of all events contained in eps_j. For example, in the first encoded period segment, since eps1 includes the four events, <*a*>, <*b*>, <*c*>, and <*d*>, and their weighted values are 0.3, 0.8, 0.9, and 0.5, respectively, $mpsw_1 = 0.9$. Besides, the weighted support upper-bound $wsub_{<ctp>}$ is the sum of the maximum period segment weight values of all encoded period segments including *ctp* over the total number of encoded period segments in *S*.

$$wsub_{<ctp>} = \frac{\Sigma mpsw_i, (<ctp>)}{m},$$

A candidate tuple pattern *ctp* is called a (frequent) weighted upper-bound tuple pattern (abbreviated as *WUBTP*) if $wsub_{ctp} \geqq min\_wsup$. If a pattern is a WUBTP, it will be checked to see if it is also a WPPP. According to definition, a *WPPP* is the one with its weighted support value $wsup_{<ctp>}$ *is* larger than or equal to *min_wsup*. If $wsup_{<ctp>} \geqq min\_wsup$, it is a weighted partial periodic pattern.

Based on the above definitions, the problem to be solved in the paper is to find all weighted partial periodic patterns, which their actual weighted support values are larger than or equal to a predefined minimum weighted support threshold, from an event sequence.

6.4 The Proposed Algorithm

In this section, the details of the proposed projection-based weighted algorithm (abbreviated as *PWA*) for mining weighted partial periodic patterns with a specific period length are described below.

6.4.1 The Proposed Projection-Based Weighted Algorithm, PWA

INPUT: An event sequence S with n events, each with a weight value, a periodic length of interest l, a minimum weighted support threshold *min_wsup*

OUTPUT: A final set of (frequent) weighted partial periodic patterns with the period length of l, *WPPPs*

STEP 1: Divide the event sequence S with the period length of interest l as multiple period segments, $S = < ps_1, ps_2, \ldots, ps_m >$. For each period segment *ps*, do the following substeps:

(a) Encode each event in ps_j to an event tuple et_{jk}, with the event and its position identifier in ps_j, and then generate the corresponding encoded period segment eps_j.
(b) Find the maximum period segment weight value $mpsw_j$ of eps_j.
(c) Put the encoded period segment with the maximum period segment weight value $mpsw_j$ in the set of encoded period segments (*EPSD*).

STEP 2: Find the weighted support upper-bound of each possible tuple *1*-pattern in the set of *EPSD*:

$$wsub_{<ctp>} = \frac{\Sigma mpsw_{j,<ctp>}}{|m|},$$

where $\Sigma mpsw_{j,<ctp>}$ is the sum of the maximum weight values of all encoded period segments including the tuple *1*-pattern (*ctp)*, and $|m|$ is the number of all encoded period segments in the set of *EPSD*

STEP 3: For each possible tuple *1*-pattern, check whether its weighted upper-bound support value is larger than or equal to the weighted minimum support threshold *min_wsup*. If yes, put the *1*-pattern in the set of weighted upper-bound tuple *1*-patterns, $WUBTP_1$; otherwise, omit the *1*-pattern

STEP 4: Gather the event tuples appeared in the set of $WUBTP_1$

STEP 5: For each segment eps_j in the set of *EPSD*, do the following substeps

(a) Get each k-th event tuple $< et_{jk} >$ in eps_j.
(b) Check whether or not the event tuple $< et_{jk} >$ appears in $WUBTP_1$. If yes, keep the $< et_{jk} >$ in eps_j; otherwise, remove the event tuple $< et_{jk} >$ from eps_j.

STEP 6: Set $r = 1$, where r represents the number of event tuples in the processed candidate patterns

STEP 7: For each pattern *tp* in the set of $WUBTP_1$, do the following substeps

(a) Find the relevant encoded period segments including X from the set of *EPSD*, and put the segments in the set of projected encoded period segments $epsd_{tp}$ of the pattern *tp*.
(b) Re-calculate the maximum weight of each projected period segment in the set of $epsd_{tp}$.
(c) Find all the weighted tuple patterns with *tp* as their prefix patterns by the *Find-WTP(tp,* $epsd_{tp}$*, r)* procedure. Let the set of returned weighted tuple patterns be WTP_{tp}.

STEP 8: Output the set of all *WPPPs.* Find all weighted tuple patterns with X' as their prefix patterns by the *Find-WTP*(X, $epsd_X$, r) procedure, where $X = X$' and $epsd_X = epsd_{X'}$. Gather the weighted upper-bound event tuple patterns X appearing in the set of $WUBTP_r$,

(a) find the weighted support value $wsup_X = \frac{avgw_X * k}{|m|}$; where $avgw_X$ is average weight value of pattern X, k is the count of encoded period segments including X, and m is the total number of encoded period segments in S.
(b) For each event tuple pattern in $WUBTP_x$, if its weighted support value $wsup_X$ is larger than or equal to the minimum weighted support threshold *min_wsup*, put the pattern in the set of weighted frequent tuple patterns, WTP_X.
(c) Return the weighted tuple patterns in the set of WTP_X.

6.5 Experimental Evaluation

6.5.1 *Evaluation of Number of Pattern and Number of Candidates*

A series of experiments was conducted to compare the performance of the traditional partial periodic patter mining algorithm, Apriori-based weighted partial periodic pattern mining algorithm (abbreviated as *AWA*), and the proposed projection-based weighted partial periodic pattern algorithm (named *PWA*) under different parameter values. The experiments were implemented in Visual 2010 C# and executed on a PC with 2.4 GHz CPU and 2.98 GB memory. To show evaluate the practical performance of the algorithms, we used IBM synthetic dataset [6] as our experimental datasets, and randomly generated the weighted values (between 0 and 1) for the 1,000 items. To build up our experimental time series dataset, each element will be selected sequentially and comma separated. For the elements in the same time point will be treated as element wrapped with brackets.

Experiments were first made on the synthetic datasets to evaluate the difference in the number of patterns obtained using *PPA* (without weight-based) [9] and *PWA* (with weight-based) while period length set at 5, 7 and 10 along with the minimum support threshold (*min_wsup*) varying from 0.05 % down to 0.01 %. Figure 6.1 reveals that *PWA* efficiently applied upper-bound weighted model to select the interested weighted candidates. And, it accelerates the mining to find meaningful WPPPs. The figure shows that the efficiency of *PWA*, especially when the minimum support threshold was decreased, or the period length increased.

Figure 6.2 showed the comparisons in the number of candidate event tuples generated by *PWA* and *AWA* algorithms for the synthetic datasets with various parameter settings, including *min_sup* and *period length (PL)*, respectively.

It could be seen in Fig. 6.2 that the numbers of weighted candidate event tuple generated by the proposed *PWA* algorithm were less than that generated by *AWA* algorithm. The main reason for this was that our proposed *PWA* algorithm reduces

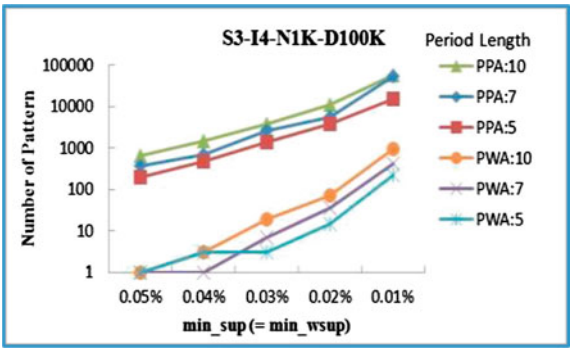


Fig. 6.1 Number of pattern for period length set at 10, 7, and 5 obtained using PPA and PWA along with different *min_wsup* values

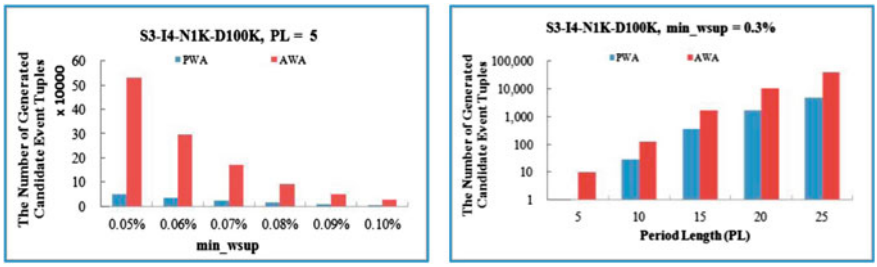


Fig. 6.2 The comparison of the numbers of candidate event tuple generated by PWA and AWA algorithms along with various parameter settings

candidate event tuples by using upper-bound weighted model. Hence, the numbers of candidates generated by *PWA* algorithm were thus obviously less than that by *AWA* algorithm.

6.6 Conclusion

The paper introduces the new research issue, named two-phased weighted partial periodic pattern mining, which considers the individual significance weights of events. In addition, a projection-based weighted upper-bound partial periodic pattern mining algorithm is proposed to find such weight-based patterns within a period length. To speed up the execution efficiency, the pruning strategy that considers the frequent relationship of events is also proposed to effectively reduce candidates in mining. The experimental results show that the memory requirement is less than that needed by the other algorithms to discover weighted partial periodic patterns.

References

1. Agrawal R, Srikant R (1995) Mining sequential pattern. In: Proceedings of the international conference on data engineering, pp 3–14
2. Ahmed CF, Tanbeer SK, Jeong BS, Lee YK, Choi HJ (2012) Single-pass incremental and interactive mining for weighted frequent patterns. Expert Syst Appl 39:7976–7994
3. Cai CH, Fu AW, Cheng CH, Kwong WW (1998) Mining association rules with weighted items. In: Proceedings of the international database engineering and applications symposium, IDEAS 98, U.K., pp. 68–77
4. Cheng H, Yan X, Han J (2004) Incspan: incremental mining of sequential patterns In iarge Databases. *SIGKDD'04*, pp. 527–532
5. Han J, Dong G, Yin Y (1999) Efficient mining of partial periodic patterns in time series databases. In Proceedings of the 15th international conference data engineering, pp. 106–115

6. IBM Quest Data Mining Project () Quest synthetic data generation code," Available via (http://www.almaden.ibm.com/cs/quest/syndata.html)
7. Liu Y, Feng J Yin J (2007) The incremental mining of constrained cube gradients. Int J Inf Technol Decis Making 6(2):253–278
8. Pinto H, Han J, Pei J, Wang K (2001) Multi-dimensional sequence pattern mining. In: Proceedings of the international conference on information and knowledge management
9. Yang KJ, Hong TP, Chen YM, Lan GC (2013) Projection-based partial periodic pattern mining for event sequences. Expert Syst Appl, Appl 40(10):4232–4240
10. Yun U, Leggett JJ (2006) WSpan: weighted sequential pattern mining in large sequence databases. In: Proceedings of the 3rd international IEEE conference on intelligent systems
11. Yun U, Ryu KH (2010) Discovering important sequential patterns with length-decreasing weighted support constraints. Int J Inf Technol Decis Making 9(4):575–599
12. Yun U, Ryu KH (2011) Approximate weighted frequent pattern mining with/without noisy environments. Knowl-Based Syst 24:73–82

Chapter 7
Edge Selection for Degree Anonymization on K Shortest Paths

Shyue-Liang Wang, Ching-Chuan Shih, I-Hsien Ting and Tzung-Pei Hong

Abstract Privacy preserving network publishing has been studied extensively in recent years. Although more works have adopted un-weighted graphs to model network relationships, weighted graph modeling can provide deeper analysis of the degree of relationships. Previous works on weighted graph privacy have concentrated on preserving the shortest path characteristic between pairs of vertices. Two common types of privacy have been proposed. One type of privacy tried to add random noise edge weights to the graph but still maintain the same shortest path. The other privacy, *k-shortest path privacy*, minimally perturbed edge weights so that there exist k shortest paths. However, the *k-shortest path privacy* did not consider degree attacks on the nodes of anonymized shortest paths. For example, if the adversary possesses background knowledge of node degrees on the shortest path, the true shortest path can be identified. We have previously presented a new concept called *(k_1, k_2)-shortest path privacy* to prevent such privacy breach [1]. A published network graph with *(k_1, k_2)-shortest path privacy* has at least k_1 indistinguishable shortest paths between the source and destination vertices. In addition, for the non-overlapping vertices on the k_1 shortest paths, there exist at least k_2 vertices with same node degree and lie on more than one shortest path. In this work, we further propose edge insertion and edge weight determination techniques to effectively achieve the proposed privacy. Numerical comparisons based on average clustering coefficient and average shortest path length show that the proposed *TNF* approach is simple and effective.

Keywords Social networks · Privacy preserving · Edge weights · K-shortest path privacy · (K_1, K_2)-shortest path privacy

S.-L. Wang (✉) · C.-C. Shih · I.-H. Ting
Department of Information Management, National University of Kaohsiung, Kaohsiung 81148, Taiwan
e-mail: slwang@nuk.edu.tw

T.-P. Hong
Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 81148, Taiwan

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_7,

7.1 Introduction

Social networking and cloud computing have become extremely popular for information sharing and knowledge management. The ease of use and free of use of current on-line social networking websites not only make sharing information so popular, simple and easy, but also expose personal identifiable information such as full name, email address, phone numbers, as well as hobbies, interests and information collected through cookies or other types of tracking mechanisms—when they are tied to individually identifiable information, to public. In addition, with the push of cloud technology, it makes on-line social networking more efficient, convenient and attracts even more users and interactions, which leads to more personal information exposable to adversaries. As such, personal attacks, reputational, financial, or family losses might occur once this personal and sensitive information falls into the hands of malicious hackers. In order to protect the privacy of users against different types of attacks, information should be anonymized before they are published.

Many practices to protect user privacy from published data have been proposed, including removing all identifiable personal information such as names and social security numbers, limiting access, "fuzzing" the data, eliminating unnecessary groupings, augmenting with additional data, etc. However, it is still easy for an attacker to identify the target by performing different structural and non-structural queries.

Recent studies show that there are four common types of privacy breaches on privacy preserving social network (graph) publishing: identity disclosure, link disclosure, and attribute disclosure. Identity disclosure refers to the threat of re-identification of nodes in a graph. For example, user profiles (such as photos, birth date, residence, interests and friend links) can be used to estimate personal identity information such as social security number. Link disclosure refers to the threat of identifying the relationship between nodes, for example, identifying the friendship between *Ada* and *Bob*. Attribute disclosure refers to the threat of identifying the attributes of a node or a link.

Privacy preservation techniques on network publishing can be classified into the followings: *k-anonymity*, *generalization*, *randomization*, *output perturbation*, and other works such as *edge weight anonymization* [2–8]. For *k-anonymity* approach, there are some studies such as *k-degree, k-neighbrohood, k-automorphism, k-symmetric, k-isomorphism, k-security*, and *k-obfuscation* privacy models. The generalization approach either generalizes nodes into super nodes or edges into super edges. One randomization approach randomly adds k false edges and deletes k true edges by keeping the number of edges unchanged. Another randomization randomly switches a pair of edges and repeats it for k times with node degree unchanged. The output perturbation approach such as differential privacy tries to propose mechanism of perturbing a data set such that the difference between the probabilities of original and perturbed (different by at most one element) data sets is within ε-differential. To protect edge weight privacy, perturbation-based

approaches to preserve linear property such as shortest paths by anonymizing the edge weights have been proposed recently [4, 9, 10].

For edge weight anonymization, current works concentrate on preserving the shortest paths characteristic between pairs of vertices [4, 10] and *k-anonymous weight privacy* [9]. To preserve the shortest paths between pairs of vertices, Gaussian randomization perturbation and greedy perturbation techniques that minimally modify the edge weights without adding or deleting any vertices and edges have been proposed [10]. The anonymized graph thus preserves the shortest path but all edge weights are perturbed. A linear programming abstract model that can preserve linear properties of edge weights (including shortest paths) after anonymization is presented in [4]. For a different type of privacy, to eliminate the distinguishability between edge weights, the *k-anonymous weight privacy* is defined in [9] as: the edge $(i \rightarrow j)$ is k-anonymous if and only if there exist at least k edges in $\Phi(i)$ whose weights $w_{i,tl}$, $l = 1,\ldots, c$, and $c \geq k$, satisfy $\| w_{i,j} - w_{i,tl} \| \leqq \mu$, $l = 1,\ldots, c$. Here, μ is a predefined positive parameter to control the degree of privacy and $\Phi(i)$ is the adjacent edge set in which all edges come from the i-th vertex.

In [6], we have proposed *k-shortest path privacy*, which minimally perturbed edge weights so that there exist k shortest paths. However, the *k-shortest path privacy* did not consider degree attacks on the anonymized shortest paths. For example, if the adversary possesses background knowledge of node degrees on the shortest path, the true shortest path can be identified. Previously we presented a new concept called *(k_1, k_2)-shortest path privacy* to prevent such privacy breach [1]. A published network graph with *(k_1, k_2)-shortest path privacy* has at least k_1 indistinguishable shortest paths between the source and destination vertices. In addition, for the non-overlapping vertices on the k_1 shortest paths, there exist at least k_2 vertices with same node degree and lie on more than one shortest path. In this work, we further propose edge insertion and edge weight determination techniques to effectively achieve the proposed privacy.

The rest of the paper is organized as follows. Section 7.2 gives the problem description. Section 7.3 describes the proposed algorithms. Section 7.4 reports the numerical experiments. Section 7.5 concludes the paper.

7.2 Problem Description

Weighted graphs can be used for analyzing the formation of communities within the network, business transaction networks, viral and targeted marketing and advertising, modeling the structure and dynamics such as opinion formation, and for analysis of the network for maximizing the spread of information through the social links [11]. Depending on the applications, the edge weights could be used to represent “degree of friendship”, “trustworthiness”, and “business transaction”, etc. In order to protect the privacy of these sensitive information (sensitive edges), two common types of privacy have been proposed. One type of privacy tried to

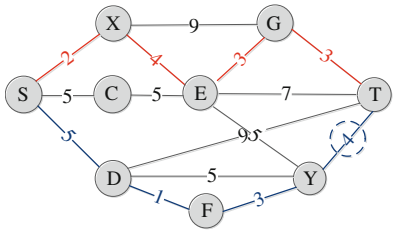


Fig. 7.1 An un-anonymized graph

add random noise edge weights to the graph but still maintain the same shortest path. The other privacy, *k-shortest path privacy*, minimally perturbed edge weights so that there exist k shortest paths. Figure 7.1 shows an undirected weighted graph with nine vertices. Assuming the relationship represented by the shortest path between vertices v_S and v_T is sensitive, $\{v_S, v_X, v_E, v_G, v_T\}$, and expected to be hidden. One possible technique is to perturb minimal number of edge weights so that there will be k shortest paths between the two vertices. Figure 7.2 shows the anonymized graph with two shortest paths $\{\{v_S, v_X, v_E, v_G, v_T\}, \{v_S, v_D, v_F, v_Y, v_T\}\}$, where the weight of edge $e_{Y,T}$ is modified to three. Therefore, given a graph G, a set of source and destination nodes H, and privacy level k, the objective of *k-shortest path privacy* is to minimally modify the graph such that there exists k shortest paths between each given pair of nodes specified in H, without adding or deleting any vertices or edges.

However, the *k-shortest path privacy* did not consider degree attacks on the nodes of anonymized shortest paths. For example, if the adversary possesses background knowledge of node degrees on the shortest path, the true shortest path can be identified. In Fig. 7.2, if the adversary knew that there is a node with degree five on the shortest path (which is vertex v_E), then the true shortest path can be inferred as $\{v_S, v_X, v_E, v_G, v_T\}$.

In order to protect the *k-shortest path privacy* from degree attacks on the anonymized paths, we presented a new concept called *(k_1, k_2)-shortest path privacy* in which a published network graph has at least k_1 indistinguishable shortest paths between the source and destination vertices [1]. In addition, for the non-overlapping vertices on the k_1 shortest paths, there exist at least k_2 vertices with same node degree and lie on more than one shortest path.

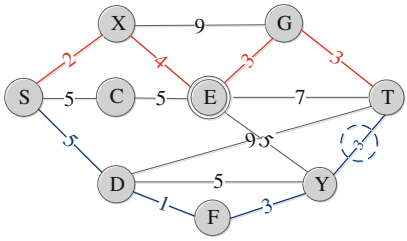


Fig. 7.2 A 2-shortest path privacy graph

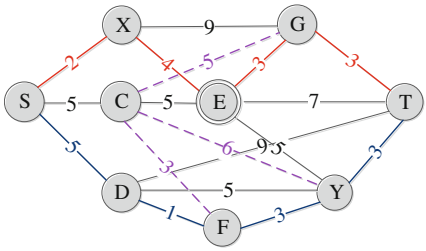


Fig. 7.3 A (2, 2)-shortest path privacy graph (by K-means)

For example, given an un-anonymized graph shown in Fig. 7.1, for $k_1 = k_2 = 2$, if edge weight $e_{Y,T}$ is modified to three, then there will be two shortest paths $\{\{v_S, v_X, v_E, v_G, v_T\}, \{v_S, v_D, v_F, v_Y, v_T\}\}$, both with path length 12, as shown in Fig. 7.3. In addition, if new edges with dash lines are added, then the *transfer nodes* on the two shortest paths will have the same node degree for at least of two nodes ($k_2 = 2$), in which the transfer nodes are the nodes on the anonymized shortest paths excluding the source, destination, and overlapping nodes.

As such, a graph with *(k_1, k_2)-shortest path privacy* will have at least k_1 indistinguishable shortest paths between the source and destination vertices. In addition, for each vertex on the anonymized shortest paths, there exist at least k_2 indistinguishable vertices with the same node degree. Therefore, given a graph G, a set of source and destination nodes H, and privacy level *(k_1, k_2)*, the objective of *(k_1, k_2)-shortest path privacy* is to minimally modify the graph such that there exists k_1 shortest paths between each given pair of nodes specified in H, and at least k_2 indistinguishable vertices with the same node degree on the k_1 shortest paths.

7.3 Proposed Algorithms

For a given pair of source and destination vertices, the objective of *(k_1, k_2)-shortest path privacy* is to modify the graph so that (1) k_1 shortest paths can be achieved, and (2) for each vertex on the same shortest path, there exists at least k_2 indistinguishable vertices with the same node degree. In this work, we propose a two-phase approach with an edge insertion and edge weight determination scheme to achieve *(k_1, k_2)-shortest path privacy*. The first phase is to modify the edge weights of the top k_1 shortest paths so that all possess the same length. The second phase is to find groups of k_2 nodes that are not on the same k_1 shortest paths, and adding edges to these nodes so that they will have same node degrees. For the first phase of achieving k_1 shortest path, we adopt the greedy approach proposed in [6]. For the second phase of finding groups of k_2 nodes, we proposed using three simple clustering strategies, namely, *modified K-means, sorting, and dynamic programming* [1]. After groups of k_2 nodes (some groups may contain more than k_2 nodes) have been obtained in previous work [1], we randomly added edges to these nodes by connecting them to nodes not on any of the shortest paths. For example,

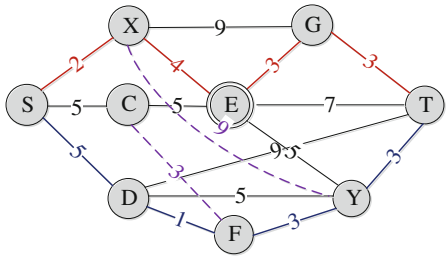


Fig. 7.4 A (2, 2)-shortest path graph (by TNF)

Fig. 7.2 shows a *2*-anonymized graph with two shortest paths from source vertex v_S to destination vertex v_T. Figure 7.3 shows the anonymized (2, 2)-*shortest path privacy* graphs using *modified K-means* clustering technique. Noted that the clusters generated are $\{(V_B, V_F), (V_G, V_D), (V_E, V_H)\}$ Fig. 7.4.

In this work, instead of connecting edges to random vertices not on the shortest paths, we propose a heuristic approach of connecting edges to transfer nodes first. When there is not enough transfer nodes, we connect the edges to vertices with highest node degree that are not on the shortest paths. The proposed modified k-means clustering algorithm and edge insertion algorithm are given as follows. The following notations will be used: *SP*, shortest path; *Nsp*, vertices on shortest paths; *TN*, transfer nodes on shortest paths, i.e., the vertices on all shortest paths excluding the source, destination, and overlapped vertices.

Input: a given graph G, source vertex V_S, destination vertex V_T, (k_1, k_2),
Output: K Clusters of vertices that vertices in the same cluster should be anonymized with same node degree,

Modified K-means-based Clustering Algorithm

1 Find top k_1 *SPs* and all N_{sp};
2 Let $TN = \{t_i \mid \text{the set of all transfer nodes}\}$;
3 If $|TN| < 2k_2$, then only one cluster;
4 Let $K = \left\lfloor \frac{|TN|}{k_2} \right\rfloor$;
5 Randomly pick K nodes as initial centroids
6 Repeat

6.1 Form K clusters by assigning each point to the closest centroid
6.2 Reassign points such that each cluster has size $k_2 \leq |C_i| \leq 2k_2 - 1$
6.3 Recompute the centroid of each cluster until centroid do not change

7 If (a cluster contains nodes all from the same SP), then
Select a vertex that has closest node degree with neighboring cluster vertex, not on the same SP, and swap the two vertices;

Transfer Node First (*TNF*) Edge Insertion Algorithm
Input: $C_1, C_2, \ldots, C_k, |C_i| \geq k_2, 1 \leq i \leq K$

Output: all vertices in the same cluster have same node degree.

1. For each cluster
   1.1 For each vertex
   1.1.1 Calculate the degree required (*degree* +) to each the maximum degree in the cluster;
2. Let $NOV = \{$the list of all vertices with non-zero node degree$\}$;
3. Sort *NOV* list in ascending order according to node degree;
4. While ($NOV \neq \phi$)
   4.1 For (the first vertex in *NOV*)
   4.1.1 While (there exits next vertex in *NOV*)
   If (there is no edge between the two nodes)
   {add an edge between them;
   decrease (*degree* +) of *N* and current vertices by one;}
   else
   Continue on next vertex in *NOV*;
   4.1.2 If ((*degree* +) of vertex $N > 0$)
   Randomly select (*degree* +) vertices not on any of the shortest paths and add edges;
   4.2 Remove first vertex from *NOV*;

In addition, the weights for the newly added edges can be determined by the following lemma.

Lemma 1 Edge Weight Insertion

Assuming a new edge $e_{X,Y}$is added between vertices *X* and *Y*

Let *S* be the source vertex and *T* be the destination vertex. The weight of $e_{X,Y}$ can be chosen by, $\left|e_{X,Y}\right| > \max\left(\left|e^1_{X,Y}\right|, \left|e^2_{X,Y}\right|\right)$ where $\left|e^1_{X,Y}\right| > SP(XT)–SP(YT)$ and $\left|e^2_{X,Y}\right| > SP(SX)–SP(SY)$.

<Pf> When a new edge $e_{X,Y}$ is added between *X* and *Y*, then shortest path between *X* and destination node *T* should remain the shortest. i.e., $\left|e^1_{X,Y}\right| + SP(Y,T) > SP(XT)$. In addition, the shortest path between source node *S* and *X* should remain the shortest, i.e., $\left|e^2_{X,Y}\right| + SP(SY) > SP(SX)$. To satisfy both condition, the weight of new edge must be greater them maximum of $\left|e^1_{X,Y}\right|$ and $\left|e^{12}_{X,Y}\right|$.

7.4 Numerical Experiments

For numerical experiments, we run simulations on a real world data set, Cora, collected in [12] and can be downloaded from LINQS webpage. The Cora dataset consists of 2,708 scientific publications classified into one of seven classes. The citation network consists of 5,429 links. Each publication in the dataset is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of 1,433 unique words. For our experiments, we randomly generate weights in the range of [1, 100] and assign to the edges.

All experiments reported in this section were performed on an Intel Pentium CPU G870, 3.10 GHz machine with 4 GB main memory, running Microsoft Windows 7 operating system. All the methods were implemented using Python programming language.

In order to examine the feasibility and characteristics of the proposed approaches, we examine the number of edges added in order to achieve the *(k_1, k_2)-shortest path privacy*. We also compare the average clustering coefficient and average shortest path length. In social networking, a graph is considered *small-world*, if (1) its average clustering coefficient is significantly higher than a random graph constructed on the same vertex set, and (2) its average shortest path length is approximately the same as its corresponding random graph. The average clustering coefficient and average shortest path length of anonymized graph are compared to the original graph to examine the characteristic that small-world property has been distorted.

For $k_1 = 2$, $2 <= k_2 <= 6$, Fig. 7.5 shows the preliminary results of number of average edges added to achieve *(k_1, k_2)-shortest path privacy* for the three proposed algorithms in [1]. It can be observed that when k_2 increases, the number of added edges increases. In addition, simple sorting based approach requires adding relatively few edges compared to two other approaches.

For $k_1 = 2$, $2 <= k_2 <= 6$, Fig. 7.6 and 7.7 show the average clustering coefficients and average shortest path lengths for the proposed algorithms. In Fig. 7.6, it can be observed that the average clustering coefficients of the *TNF* edge insertion approach is much closer to the average clustering coefficients of original graph, compared to the random insertion approach, for all three vertex clustering algorithms proposed in [1]. In Fig. 7.7a, the average shortest path lengths of

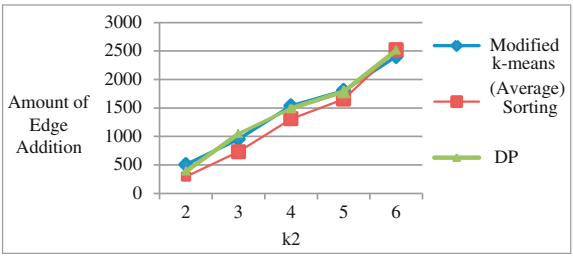


Fig. 7.5 Number of edges added

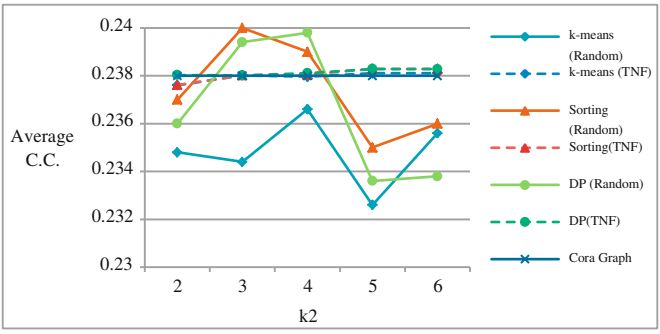


Fig. 7.6 Average clustering coefficients

random insertion approach decrease as k_2 increases. However, in Fig. 7.7b, the average shortest path lengths of *TNF* approach remain very close to the original graph (scale in two figures are slightly different). Overall, the proposed *TNF* edge insertion approach incurs less modification on the graph characteristics.

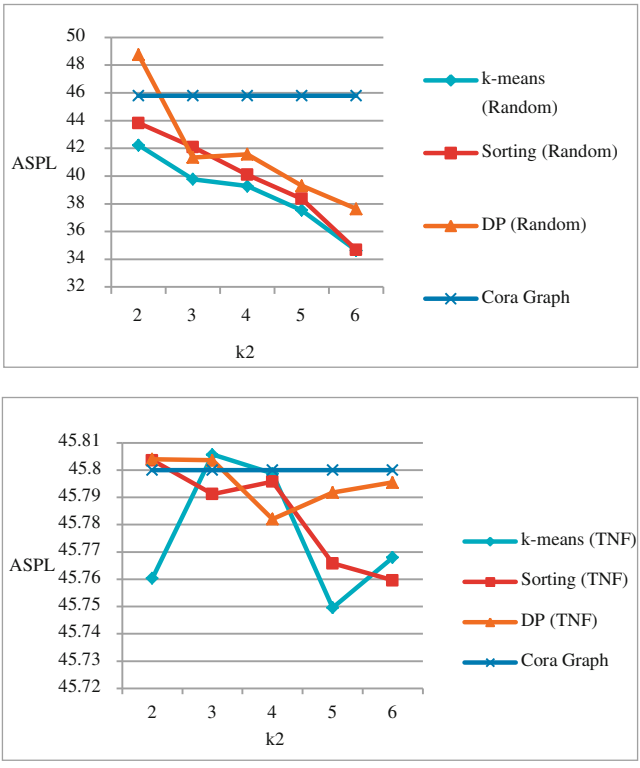


Fig. 7.7 **a** Average shortest path lengths (by random) **b** Average shortest path lengths (by TNF)

7.5 Conclusions

To protect the *k-shortest path privacy* from degree attacks on the nodes of anonymized paths, we have presented a new concept called *(k_1, k_2)-shortest path privacy* in which a published graph has at least k_1 indistinguishable shortest paths between the source and destination vertices. In addition, for the non-overlapping vertices on the k_1 shortest paths, there exist at least k_2 vertices with same node degree and lie on more than one shortest path [1]. In this work, we further propose edge insertion and edge weight determination techniques to achieve the proposed privacy. Numerical comparisons based on average clustering coefficient and average shortest path length show that the *TNF* approach is simple and effective. However more works need to be done. We plan to investigate different ways of adding edges, distributing k_2 nodes evenly among k_1 shortest paths, and experiment on larger data sets for more conclusive results.

Acknowledgments This work was supported in part by the National Science Council, Taiwan, under grant NSC 101–2221-E-390–028-MY3.

References

1. Wang SL, Shih CC, Ting HH, Hong TP (2013) Degree anonymization for k-shortest-path privacy, submitted to 2013. IEEE international conference on SMC, Manchester, October 2013
2. Government Information Laws. www.privacyinternational.org/foi/survey
3. Cheng J, Fu A, Liu J (2010) K-isomorphism: privacy preserving network publication against structural attacks. In: SIGMOD conference, 459–470
4. Das S, Egecioglu O, Abbadi AE (2010) Anonymizing weighted social network graphs. In: ICDE, 904–907
5. Wang SL, Tsai YC, Kao HY, Hong TP (2010) Anonymizing set-valued social data. The 2010 international symposium on social computing and networking (SocialNet'10), Hangzhou, December 2010
6. Wang SL, Tsai ZZ, Hong TP, Ting HH (2011) Anonymizing shortest paths on social network graphs. The 3rd Asian conference on intelligent information and database systems (ACIIDS), Daegu, April 2011
7. Zhou B, Pei J (2008) Preserving privacy in social networks against neighborhood attacks. In: ICDE, 506–515
8. Zou L, Chen L, Ozsu MT (2009) K-automorphism: A general framework for privacy preserving network publication. In VLDB, 200
9. Liu L, Liu J, Zhang J (2010) Privacy preservation of affinities in social networks. In: ICIS
10. Liu L, Wang J, Liu J, Zhang J (2009) Privacy preservation in social networks with sensitive edge weights. In: SDM, 954–965
11. LINQS, Statistical relational learning group at University of Maryland, USA, http://www.cs.umd.edu/projects/linqs/projects/lbc/
12. Liu K, Terzi E (2008) Towards identity anonymization on graphs. In: SIGMOD Conference, 93–106

Chapter 8
K-Neighborhood Shortest Path Privacy in the Cloud

Shyue-Liang Wang, Jia-Wei Chen, I-Hsien Ting and Tzung-Pei Hong

Abstract Preserving privacy on various forms of published data has been studied extensively in recent years. In particular, shortest distance computing in the cloud, while maintaining neighborhood privacy, attracts latest attention. To preserve fixed-pattern one-neighborhood privacy, current approach requires the calculation of all-pairs shortest paths in advance, which is time consuming for large graphs. In this work, we propose a new flexible k-neighborhood privacy-protected and efficient shortest distance computation scheme in the cloud. Combining k-skip shortest path sub-graphs, vertex hierarchy labeling and bottom-up partitioning, the proposed technique not only subsumes one-neighborhood privacy but also provides efficient partitioning and query processing. Numerical experiments demonstrating the characteristics of proposed approach are presented.

Keywords Privacy preservation · *k*-neighborhood privacy · Shortest path distance · *k-skip*

8.1 Introduction

Many real-world applications can be modeled as weighted and/or un-weighted graphs, e.g., social networking, web applications, transportation, communication, and business transaction networks, etc. But operations on graphs are usually time consuming due to structural complexity and large graph sizes. The rise of cloud

S.-L. Wang (✉) · J.-W. Chen · I.-H. Ting
Department of Information Management, National University of Kaohsiung, Kaohsiung, Taiwan
e-mail: slwang@nuk.edu.tw

T.-P. Hong
Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 81148, Taiwan

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_8,

computing provides a new platform for storage and computation intensive operations, such as graph operations. Cloud computing creates a fluid pool of resources across servers and data centers that enable (client) users to access (outsourced) stored data and applications on an as-needed basis. It is therefore desirable to employ cloud computing to manage large graphs and process complex operations efficiently. However, one of the concerns using these services is the negative effects on privacy, e.g., the clients may be unwilling to outsource their valuable datasets as sensitive information might be exposed.

Several practices to protect user privacy from published data have been proposed, including removing all identifiable personal information such as names and social security numbers, limiting access, "fuzzing" the data, eliminating unnecessary groupings, augmenting with additional data, etc. However, it is still easy for an attacker to identify the target by performing different structural and non-structural queries.

Recent studies show that there are four common types of privacy breaches on privacy preserving social network (graph) publishing: identity disclosure, link disclosure, attribute disclosure, and edge weight disclosure. Identity disclosure refers to the threat of re-identification of nodes in a graph. For example, user profiles (such as photos, birth date, residence, interests and friend links) can be used to estimate personal identity information such as social security number. Link disclosure refers to the threat of identifying the relationship between nodes, for example, identifying the friendship between *Ada* and *Bob*. Attribute disclosure refers to the threat of identifying the attributes of a node or a link. Edge weight disclosure refers to the threat of identifying the weights of a link or a path.

For edge weight anonymization, current works include: (1) preserving the shortest paths characteristic between pairs of vertices [1, 2], (2) *k-anonymous weight privacy* [3], (3) *k-shortest path privacy* [4], and (4) neighbor privacy for shortest paths in the cloud [5]. To preserve the shortest paths between pairs of vertices, Gaussian randomization perturbation and greedy perturbation techniques that minimally modify the edge weights without adding or deleting any vertices and edges have been proposed [2]. The anonymized graph thus preserves the shortest path but all edge weights are perturbed. A linear programming abstract model that can preserve linear properties of edge weights (including shortest paths) after anonymization is presented in [1]. For a different type of privacy, to eliminate the distinguishability between edge weights, the *k-anonymous weight privacy* is defined in [3] as: the edge $(i \rightarrow j)$ is k-anonymous if and only if there exist at least k edges in $\Phi(i)$ whose weights $w_{i,tl}$, $l = 1,\ldots, c$, and $c \geq k$, satisfy $\| w_{i,j} - w_{i,tl} \| \leqq \mu$, $l = 1,\ldots, c$. Here, μ is a predefined positive parameter to control the degree of privacy and $\Phi(i)$ is the adjacent edge set in which all edges come from the i-th vertex. For *k-shortest path privacy*, several heuristic algorithms were proposed to minimally perturb edge weights so that there become k shortest paths in a graph. In addition, to deal with degree attacks on the anonymized k shortest paths, a new concept called *(k_1, k_2)-shortest path privacy* was proposed [6]. To preserve neighbor privacy for computing shortest path distance in the cloud, a graph is transformed to a link graph on client side and a number of outsourced graphs on

the cloud server side. The shortest path distance can be answered using both types of graphs together, and *one-neighborhood* privacy on the cloud server side can be preserved. A greedy based graph generation scheme was proposed [5].

In this work, we extend the *one-neighborhood* privacy concept and propose a new flexible *k-neighborhood* privacy-protected and efficient shortest distance computation scheme in the cloud. Adopting *k-skip* shortest path sub-graphs, vertex hierarchy labeling and bottom-up partitioning, the proposed technique not only subsumes *one-neighborhood* privacy but also provides efficient partitioning and query processing. Numerical experiments demonstrating the characteristics of proposed approach are presented.

The rest of the paper is organized as follows. Section 2 gives the problem description. Section 3 describes the proposed algorithms. Section 4 reports the numerical experiments. Section 5 concludes the paper.

8.2 Problem Description

For the *one-neighborhood privacy* in [5], given a fixed number of p outsourced servers, the objective is to transform a large weighted un-directed graph into p sub-graphs that are to be stored in s servers respectively, so that (1) shortest distance between any two vertices can be efficiently computed in the cloud, (2) neighborhood attacks can be prevented on each server, and (3) overhead on the client side can be minimized.

For example, given an original graph shown at the bottom of Fig. 8.1, and assuming the number of server on the cloud is $p = 2$. Fig. 8.1 shows two transformed sub graphs on the top and the link graph at the bottom. The link graph will be saved in the client side (data owner) and the two sub graphs will be stored in two servers in the cloud respectively. Note that not all links are shown here to keep the figure clear and ease to read.

For a query to calculate the shortest path distance on neighboring vertices such as A and B, vertex A is on G_1 and vertex B is on G_2. The search then has to start from original graph, e.g. vertex A. Since here are two vertices $\{B,L\}$ are graph G_2 that are linked to vertex A, we calculate the distance $d(A,B) = |\{A,B\}| = 1$ and $d'(A,B) = |\{A,L,B\}| = 1 + 2 = 3$. The shortest path distance is the minimum of the two, which is one.

For a query on non-neighboring vertices such as E and J, both vertices are not on graph G_1 or graph G_2. However, vertex E is linked to vertices C and F on graph G_1 and vertex J is linked to vertices H and K on graph G_1. The shortest path distance between E and J can be calculated as $d(E,J) = |\{E,F,H,J\}| = 1 + 4+4 = 9$.

To obtain the transformed sub graphs, the scheme proposed by Gao etc. [5] required to calculate all-pairs shortest paths first and then randomly select a set of shortest paths to form a sub graph, which is to be stored in one server. Random selection of shortest path may cause overlaps between sub graphs. Calculating all

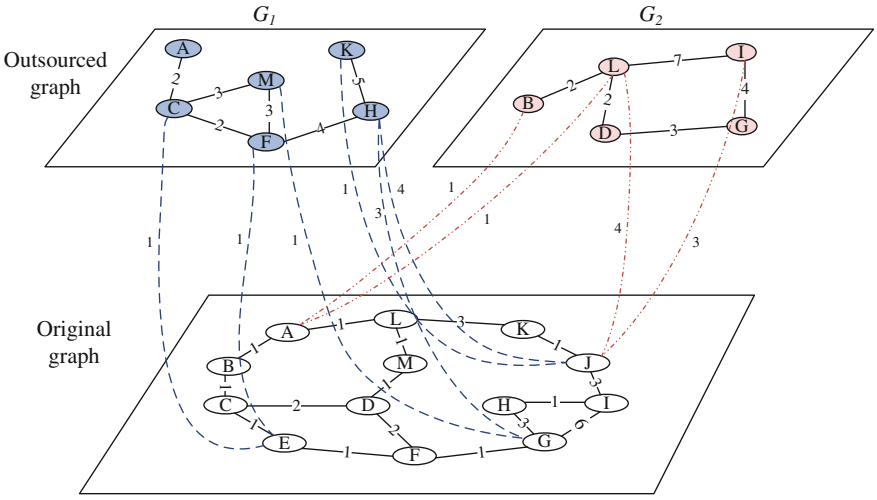


Fig. 8.1 Link graph and outsourced graphs (by Gao)

shortest path may consume a lot of computation effort. In addition, if all shortest paths have been obtained, their shortest paths and distances could be saved directly (either on servers or on client), instead of saving sub graphs which required more calculation to obtain results already known. As such, we propose a new flexible *k-neighborhood* privacy-protected scheme for efficient computation for shortest path in the cloud.

To avoid calculating all-pairs shortest path, we adopt *k-skip* shortest paths [7] and vertex labeling hierarchy [8] techniques. For example, Fig. 8.2 shows an original graph at the bottom, a *3-skip* graph in the middle and two sub graphs on the top to be stored in the cloud.

Given two vertices *x, y* in a graph, if *P* is the shortest path from *x* to *y* and *P** is a subset of the vertices in *P*, *P** is a *k*-skip shortest path from *x* to *y*, if it includes at least a vertex out of every *k* consecutive vertices in *P* [7]. In general, *P** describes *P* by sampling the vertices in *P* with a ratio of at least *1/k*.

The middle figure in Fig. 8.2 is a *3-skip* shortest path graph generated from the original graph at the bottom. All-pairs shortest path calculation is not required here. The two sub graphs on the top can be obtained by the vertex labeling and Dijkstra bi-directional partitioning techniques efficiently. Notice that the proposed *k-skip* graph is more flexible and general than *one-neighborhood* privacy graph. Given two vertices on a sub graph in the cloud server, there may be one or two hidden neighboring nodes for *one-neighborhood* privacy graph. However, for proposed *k-skip* approach, depending on *k*, there may be zero to *2*(k−1)* hidden neighbors between two vertices on a sub graph.

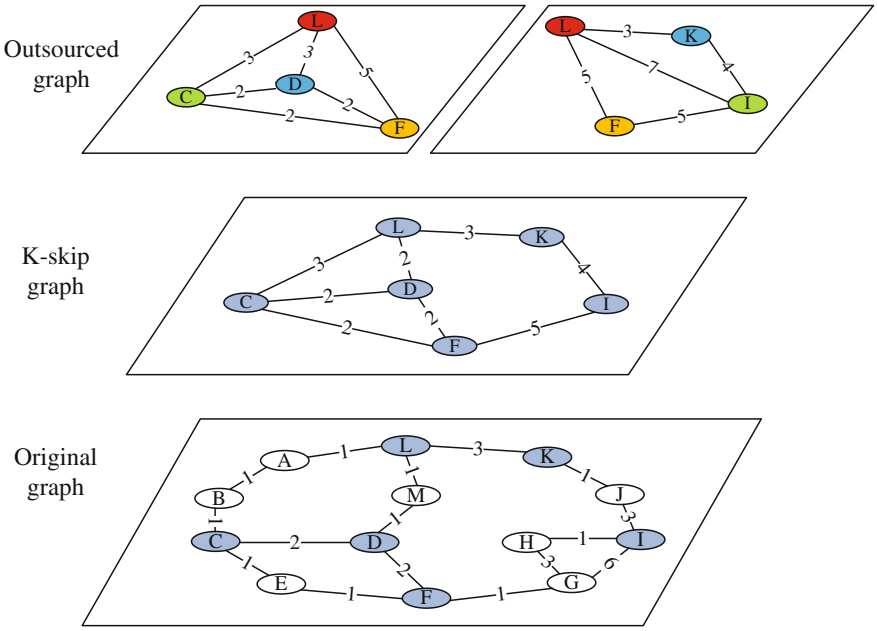
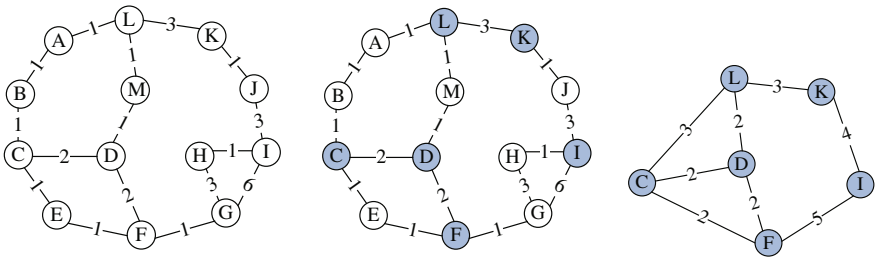


Fig. 8.2 3-*skip* graph and outsourced graphs

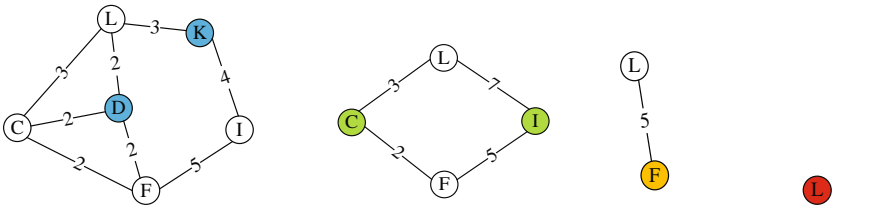
8.3 Proposed Algorithms

To achieve *k-skip* neighborhood privacy for computing shortest path distance in the cloud, we propose a three-step scheme: (1) building *k-skip* sub graph, (2) building vertex labeling hierarchy, and (3) bottom-up partitioning of sub graphs, as shown in Figs. 8.3, 8.4, and 8.5, respectively. The sub graphs built will be stored in the outsourced graphs respectively.



(a) Original graph **(b)** 3-skip cover **(c)** 3-skip graph

Fig. 8.3 *k-skip* graph



(a) $G_1 = G$, $L_1 = \{D, K\}$ **(b)** G_2, $L_2 = \{C,I\}$ **(c)** G_3, $L_3 = \{F\}$ **(d)** $L_4 = \{L\}$

Fig. 8.4 Construction of the vertex hierarchy

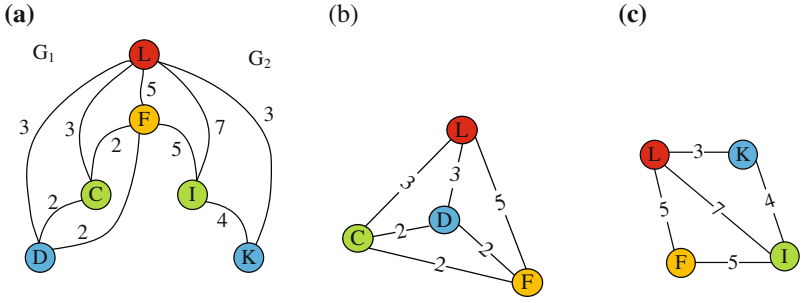


Fig. 8.5 Bottom-up partition. **a** Bottom-up Partition. **b** Outsourced graph G_1. **c** Outsourced graph G_2

For the first step, the *k-skip* graph was introduced in the context of spatial network databases [7]. It is able to answer *k-skip* queries significantly faster than finding the original shortest paths in their entirety. In addition, it carries a structure that occupies less space than storing the underlying road network. The technique first builds a *k-skip* cover, shown in Fig. 8.3b, which contains vertices with higher node degrees *{C, D, F, I, K, L}* and can reach any vertex in the graph within *(k−1)* steps on the shortest path. It then builds *k-skip* graph based on the cover vertices where the edge between two vertices represent the shortest path between them.

For the second step, vertex labeling hierarchy is to add shortest distances between levels of vertices in a hierarchical manner. The idea is to pre-calculate the shortest path distances to save query processing time later. It starts with selecting independent set *{D, K}* from graph G_1, as shown in Fig. 8.4a. The technique starts with selecting a vertex with lowest node degree in the graph, deleting its direct neighboring nodes and repeats the process. The selected vertices form an independent set. Figure 8.4b shows a reduced graph after deleting the independent set of vertices. For the given example, the graph is reconstructed into a 4-level hierarchy.

The algorithm for constructing such vertex hierarchy is given as following.

Algorithm 1 : construct Vertex Hierarchy

Input : G

Output : G_H

initialize $G_H := G_i := G$

while $(|V_H| > 0)$:

$L_i := \text{maximun independent set of } G_i$

$E' := \varphi$

for each vertex $u \in L_i$:

remove vertex u from G_i

for each edges $(u,v),(u,w) \in E_i$:

if $d(u,v) + d(u,w) < dist(v,w)$ then

$w_H(v,w) := w_i(v,w) := tmp$

else :

$E' := E' \cup \{(v,w,d(v,w))\}$

build $\forall e' \in E'$ to G_i and G_H

For the third step, based on Dijkstra's bi-directional search, the *k-skip* graph is partitioned in a bottom-up manner into *p* sub graphs to be stored in *p* cloud servers. The process starts with partitioning the level one independent set into *p* subsets. It then builds up the vertices in the second level from vertex hierarchy and repeats the process to the highest level. Figure 8.5a shows the constructed graph. Figure 8.5b, c demonstrate the final two partitions, as current example has only two vertices in the level one independent set. The algorithms for the partitioning of sub graphs are given as follows.

Algorithm2 : partitionGraph

Input : B_S, G_H

Output : G_i

initialize : bottom:= $\{\forall u \in G_H \mid level_{G_H}(u) = 1\}$

initialize : #partition, $G_P := \phi$ // number of partitions

pNum:= $\lceil |bottom| / \#partition \rceil$

while $(|bottom| > 0)$

$B_S :=$ pick #pNum of vertex $u \in$ bottom randomly, and not replace

$G_O(i) :=$ *generate Subgraph* (B_S, G_H), i = 1,...,#partition

$G_P := G_P \cup \{G_O(i)\}$

Return G_P

Algorithm3 : generate Subgraph

Input : B_S, G_H

Output : G_{sub}

initialize : $G_{sub} := empty\ graph$, $Layer := bottom$

while ($|Layer| > 0$)

upp := ϕ

for each vertex u $\in$ Layer

for $(u, v) \in E : level_{G_H}(v) > level_{G_H}(u)$

build edge (u, v) on G_{sub}

if $v \notin upp$ then

upp := upp $\cup \{v\}$

$level_{G_{sub}}(u) := levelG_H(u)$

Layer := upp

Return G_{sub}

8.4 Numerical Experiments

For numerical experiments, we run simulations on synthetic datasets generated by *Networkx* library graph generator [9]. Then *Power_law_cluster_graphs* were generated with 3,000, 6,000, and 9,000 nodes respectively. The number of random edges added to each node is set as one and the probability of adding a triangle after adding a random edge is set as 0.01. The graphs produced here follow the small-world phenomena, as shown in Fig. 8.6. In social networking, a graph is considered small-world, if (1) its average clustering coefficient is significantly higher than a random graph constructed on the same vertex set, and (2) its average shortest path length is approximately the same as its corresponding random graph. In our datasets, the numbers of edges are 2,999, 5,999, 8,999, and the average numbers of edges of all shortest paths are 7.39, 8.32, and 9.16 respectively.

To simulate the cloud environment, we build a one-master, 5-slaves distributed system. The master machine works as local client and the five slaves work as outsourced server in the cloud. The virtual machine are built on three 32-bit Pentium(R) dual-core CPU 3.20 GHz machine with 4 GB memory running windows 7 and one 32-bit Celeron(R) CPU 1.80 GHz machine with 2 GB machine running windows 7, by using *Virtual Box* [10] under Linux Ubuntu operating system. The proposed algorithms are implemented in Python and *IPython.parallel* package [11].

To examine the characteristic of *k-skip* graphs and vertex hierarchy graphs, Fig. 8.7 shows the *k-skip* construction effect and hierarchy shortcut addition effect, where $|V'|$ and $|E'|$ are number of vertices and edges respectively in *k-skip* graph and $|E^*|$ is the number of edges in vertex hierarchy graph. It can be seen that as

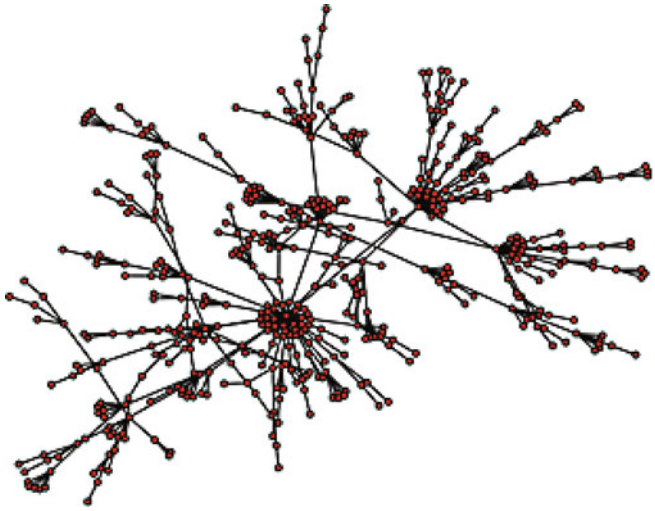


Fig. 8.6 Dataset Graph

	Original		K=2			K=3			K=4		
Dataset	|V|	|E|	|V'|	|E'|	|E*|	|V'|	|E'|	|E*|	|V'|	|E'|	|E*|
p3000	3000	2999	919	919	992	616	762	808	472	799	835
p6000	6000	5999	1821	1830	1990	1230	1563	1676	1040	1881	1957
p9000	9000	8999	2751	2762	3009	1835	2222	2428	1432	2322	2438

Fig. 8.7 *k-skip* contraction effect and hierarchy shortcut addition

k increases, the number of vertex $|V'|$ and the number of edge $|E'|$ decreases accordingly. In addition, the number of shortcut addition $|E^*|$ also decreases. However, the average node degree increases ($|E'|/|V'|$) when k increases. This implies that when k increases, the density of the contracted graph also increases. To examine the construction times, when k increases, Fig. 8.8 shows that the *k-skip* construction time increases and Fig. 8.9 shows that hierarchy graph construction time increases as well. The increase is due to more edges need to be searched for larger k. The effects indicate that it takes more computation time to construct higher k *k-skip* graphs and vertex hierarchy graphs.

To examine the query processing time, Fig. 8.10 shows that when k increases, the query time increases. The times showed are averages of fifty pairs of randomly selected sources and destination vertices on dataset p6000. The increase of query time is due to the increase of node degrees and therefore requires more time to search for shortest paths.

Figure 8.11 shows that the query time decreases when number of sub graph increases. The effects indicate that the time to calculate the shortest path distance will be shortened when number of servers in the cloud is increased up to five.

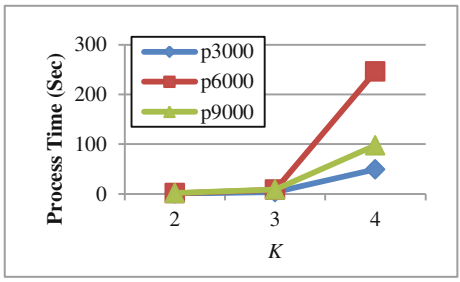


Fig. 8.8 k-sikp graph construction time

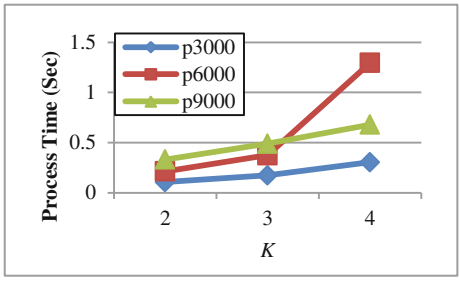


Fig. 8.9 Hierarchy graph construction time

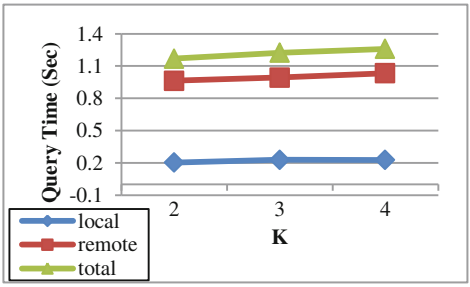


Fig. 8.10 Query time varying *k*

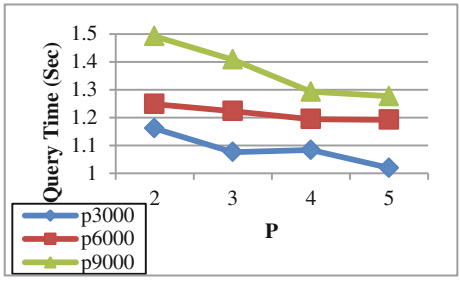


Fig. 8.11 Query time varying *p*

8.5 Conclusions

To protect neighborhood privacy and efficient calculation of shortest path distance in the cloud environment, we propose a new scheme by combining *k-skip* shortest path calculation, vertex hierarchy labeling and bottom-up partition techniques. The proposed technique is more flexible and efficient than *1-Neighborhood* privacy approach. Numerical experiments demonstrating the characteristic and efficiency of proposed approach are presented. However, more work needs to be done. We plan to investigate on different characteristic and different size of datasets and examine the effect on different number of servers, for more conclusive results.

Acknowledgments This work was supported in part by the National Science Council, Taiwan, under grant NSC 101-2221-E-390 -028 -MY3.

References

1. Agrawal D, El Abbadi A, Antony S, Das S (2010) Data management challenges in cloud computing infrastructures. In: Proceedings of the 6th international conference on databases in networked information systems, Berlin, pp 1–10
2. Das S, Egecioglu O, El Abbadi A (2010) Anonymizing weighted social network graphs. In: 2010 IEEE 26th international conference on data engineering (ICDE), pp 904–907
3. Fu AW-C, Wu H, Cheng J, Chu S, Wong RC-W (2012) IS-LABEL: an independent-set based labeling scheme for point-to-point distance querying on large graphs. arXiv:1211.2367
4. Gao J, Yu JX, Jin R, Zhou J, Wang T, Yang D (2011) Neighborhood-privacy protected shortest distance computing in cloud. In: Proceedings of the 2011 ACM SIGMOD international conference on management of data, New York, pp 409–420
5. IPython.parallel, http://ipython.org/ipython-doc/dev/parallel/
6. Liu L, Liu J, Zhang J (2010) Privacy preservation of affinities in social networks. In: ICIS
7. Liu L, Wang J, Liu J, Zhang J (2009) Privacy preservation in social networks with sensitive edge weights. In: SDM, pp 954–965
8. Wang SL, Shih CC, Ting HH, Hong TP (2013) Degree anonymization for K-shortest-path privacy. In: IEEE international conference on SMC, Manchester, (submitted)
9. Wang SL, Tsai ZZ, Hong TP, Ting HH (2011) Anonymizing shortest paths on social network graphs. In: The third asian conference on intelligent information and database systems (ACIIDS), Daegu
10. Networkx, http://networkx.github.io/
11. Tao Y, Sheng C, Pei J (2011) On k-skip shortest paths. In: Proceedings of the 2011 ACM SIGMOD international conference on management of data, New York, pp 421–432
12. VirtualBox, https://www.virtualbox.org/

Chapter 9
The Framework of Information Processing Network for Supply Chain Innovation in Big Data Era

Chian-Hsueng Chao

Abstract The challenges of the global marketplace and the growing complexity of business philosophies and technologies mix, the enterprises are forced to utilize knowledge, capabilities, and resources to be found within and outside their information processing networks. The enterprises are demanding more than just access to data, they want processed and refined big data and information to help them to reach more responsive and effective tactical decisions. Under this paradigm shift, data and information-oriented productivity depends on the sharing of knowledge and skills among workers, so that enterprise strategies can be driven by the collective intelligence and competence of the group to face business challenges and enable organizational learning and innovations. In the cloud computing and big data era, management of enterprise knowledge to create business values and competitive advantages is especially important for supply chain practices. This paper focuses on the development of enterprise information processing network and application framework that bind organizational strategies, business processes, data, information, technologies, and people together to better utilize knowledge in business practices. The ultimate goal is the transformation of an enterprise network into a knowledge network for supply chain organic innovations!

Keywords Information processing network · Knowledge management · Supply chain management · Big data analytics

C.-H. Chao (✉)
Department of Information Management, National University of Kaohsiung, 700, Kaohsiung University Rd, Nanzih District, 811, Kaohsiung, Taiwan, R. O. C
e-mail: cchao@nuk.edu.tw

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_9,

9.1 Introduction

Today, the Supply Chain Management (SCM) is a boundary-spanning, channel-unifying, dynamic, and coevolving philosophy of inter-enterprise management. The major contribution of today's supply chain model is to improve and enhancing collaboration between businesses and their trading partners [1]. In today's business practices, competing for supply chain requires the alignment of corporate strategies to what the organization knows, or developing knowledge management (KM) capabilities to support a desired supply chain solution. Management of organizational knowledge for creating business values and generating competitive advantages is critical for organizational survival. A good knowledge management should support people to access and learn from past and present organizational business practices/strategies and to apply the lessons learned when making future decisions. Therefore, a successful knowledge-oriented business for organizations should link supply chain management, relationship management, and knowledge management to function in an adaptive way to cope with every changing business challenges.

9.2 The Big Data Impacts

Big data, as the next frontier for innovation, competition, and productivity [2] will have tremendous impact on our daily life. In big data era, the flow of data can be among different devices and in different types. The data can be any type with any forms, such as business data, social network messages, blog, forum, web page, multimedia, SMS, email, sensor data (e.g. NFC, GPS, RFID, M2M), and so forth. Because the data came from a variety of sources, the big data is considered to be at the scale of up to Zettabyte. Therefore, timely and cost-effective analytics over big data is now a key ingredient for success in many businesses, scientific and engineering disciplines, and government endeavors [3]. The characteristics of big data analytics are variety, speed, and with big volume, and therefore the manipulation of data relies on intelligent approaches to deal with growing structured, semi-structured or unstructured big data. Currently, the cloud computing and parallel computing do much of the work in big data analytics.

For the past few decades, enterprises have constantly reinvented themselves through a series of business and technological innovations to fit into the global spectrum of business. The increasing adaptivity and responsiveness of business practices has led to the role of big data analytics in business practices. Big data analytics relies heavily on the interpretation of data into useful knowledge for enterprise or supply chain to make more responsive and effective tactical decisions. Information, such as demographics, consumers' behaviors, and numerous other business statistics, and the associated processing power are critical for the survival of enterprises in business. With the global deployment of computers, mobile devices, and interconnecting networks, participants can work collaboratively, share

networked resources, exchange knowledge, and improve corporate or supply chain performance. Corporate and supply chain strategies can be driven by the collective intelligence of groups to better meet today's business challenges.

9.3 The Value of Data and Information in Supply Chain

Mentioning about the values, Porter's value chain [4] concept is well adopted by organizations to profile their competitiveness and business values. The value chain divides the organization into a set of generic functional areas, which can be further divided into a series of value activities. An enterprise will be profitable as long as it creates more value than the cost of performing its value activities [5]. To model a business system, the effort for the separation of a complex part from the whole in which we are interested is called an Abstraction. This is a very practical methodology for the modelling of a complex system, especially in the supply chain modelling effort. Through abstractions, any complex business object in a system can be denoted as a black box that produces certain outputs regardless of its internal complexity. And later, when necessary, this abstract object can be further analyzed and broken down into several sub-objects. Therefore we can model supply chain business process integration based on value chain as shown in Fig. 9.1.

In Fig. 9.1, recall that Porter's value chain described an enterprise as a set of generic functional areas (such as inbound logistics, operations, marketing, outbound logistics, etc.). Porter also recognized linkages outside the enterprise, as

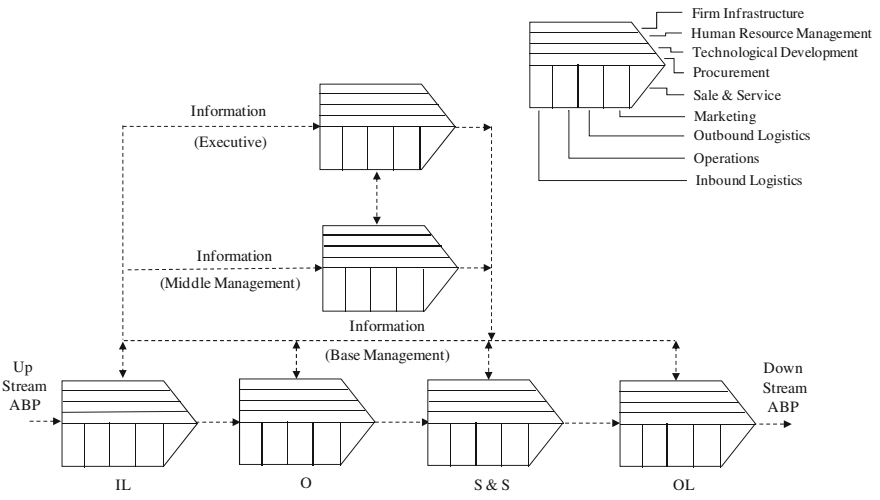


Fig. 9.1 Global network of value chain in abstract business process (after Gale and Eldred, 1996–modified)

they relate to the customer's perception of value. Therefore, it is also an open structure, and the network can be developed in a fractal pattern just like the extended value chains. To this point, a Porter value chain is an abstraction of a business process, because an enterprise is a business process entity in a global information-processing network. Therefore there is a different from the original model proposed by Gale and Eldred [6], which focused on the process view of abstraction instead of the global supply chain value management scheme.

Again, the term "abstraction" is used here to describe the generalization of any business process for the modelling purpose. Through this characterization, a business process can be generalized into what we call the Abstract Business Process (ABP). The Abstract Business Process is just like a business process which can be decomposed into several sub-processes, which is, an ABP is made up of lower level ABPs. This interconnected value chain system can act as a supply chain or information processing network that encompasses the modern business world, and participating organizations can readily extend their technologies and knowledge to their partners. The extended enterprise aspect enables supply chain integration and more effective outsourcing solutions for both internal and external stakeholders [7].

On the other hand, the rapid growth of the Internet brought about new business philosophies and fostered the growth of new strategic alliance, data, information, and business process integration across the borders of enterprises. The information processing view of an organization has been considered one of the most influential contributions to the contingency literature [8]. In this philosophy, information processing network provides the channels for exchange and processing of information in a global system. The primary role of the information processing network is to provide information exchange among its subsystem—the information processing nodes as shown in Fig. 9.2.

In Fig. 9.2, the information processing network view of "virtual enterprise" can be from different divisions, departments, or organizations. The information-processing nodes within the network are responsible for sending, receiving, selecting, producing, and communicating (i.e. exchange data and information) with other information processing nodes. An organization's value chain consists of all activities performed to design, produce, market, deliver, and support its product and service. For the analysis of business data communication, the information processing network connects its nodes, which in turn, are organized into business components. The business components of the organization include people, processes, events, machines, and information that interact and combine to produce the outputs (e.g. information, product, service) of the organization.

Recalled that a knowledge-enabled organization is a learning organization, one where all employees are using their knowledge, skills, and learning to meet today's business challenges and to create new opportunities for the future. Therefore, the value is created whenever information flows through the information process nodes and the information processing network. In business practices, collaborative problem solving, conversations, and teamwork generate a significant proportion of

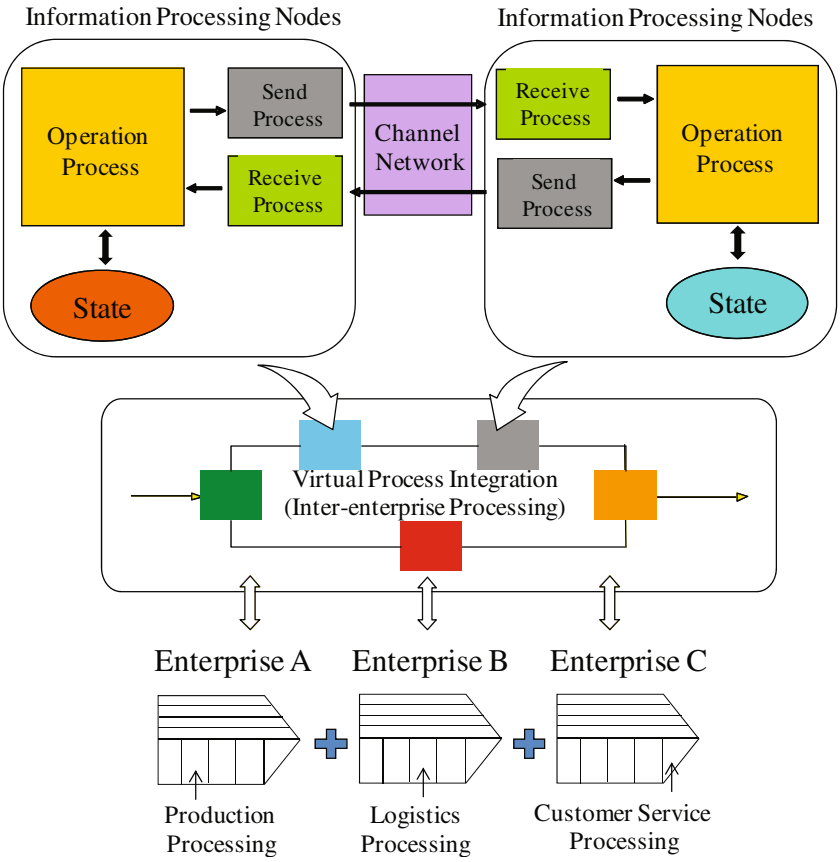


Fig. 9.2 The information processing nodes for value creation

the knowledge assets that exist within a firm or entire supply chain. With network connectivity, the virtual enterprise can work collaboratively to share knowledge and best practices that enable supply chain "co-evolving."

The beauty of a supply chain knowledge network is that the true value of the information surpasses the conventional boundaries that often restrict employees' thinking [9]. The information-processing nodes within each network in either organization can work collaboratively to achieve strategic goals in the newly joined network. Therefore, values include all information that flows through an organization and between an organization and its suppliers, its distributors, and its existing or potential customers. Indeed, data and information defines business relationships.

9.4 Portal of Storm: Streams Computing, Data Analytics and Organic Innovation

Data analytics offers values! Enterprises want analytics to exploit their growing data and computational power to get smart, get innovative that they never could before [10]. Therefore, enterprises are becoming data and knowledge intensive instead of capital intensive. In business practices, data is the basic building block of information, whereas information is the glue that unifies businesses partnerships and ultimately of a knowledge-based business.

Today, the streams computing and cloud computing are two major methodologies to deal with big data analytics. The streams computing focuses on processing big and continuous "motion" data with less than a microsecond in a real-time basis. Unlike traditional data analysis approach, the big data analytics can be processed and analyzed before being store in the database. Whereas, the cloud computing offers a variety of flexible computing schemes to deal with data in a distributed manner.

Senge [11] in his book on system thinking described a learning organization as "an organization that is continually expanding its capacity to create its future." The information processing network and data analytics made organizational learning more effective and will enable organization to behave like living organic structure to adapt and evolve in a changing environment. Interdependency between core knowledge workers will increase as the success of the enterprise becomes more dependent on how well they integrate their knowledge to produce innovative products and services [12]. The innovation is never been close to reach for each supply chain members.

9.5 The Framework for Big Data Applications

With the growing maturity of Internet technology, cloud computing, streams computing, and big data analytics, the KM system can become more active. Big data applications are needed for the entire supply chain knowledge network, because the effectiveness of an supply chain solution will depend largely on its ability to deliver an accurate and common view of customer demand data, as well as any subsequent events, plans, or other business data. Knowledge management system must be able to capture, process, refine big data and information from different data sources that employees need. The technologies to facilitate these highly interactive communications are summarized in Fig. 9.3.

In this figure, there are business application and system domain, servers, and big data analytics domain. The big data analytics domain consists of integrated knowledge application, composite repository and database. The business application and system domain provides greatest possible access for diverse computing devices, such as PC, NB, Tablet PC, Smartphone, and PDA to use enterprise

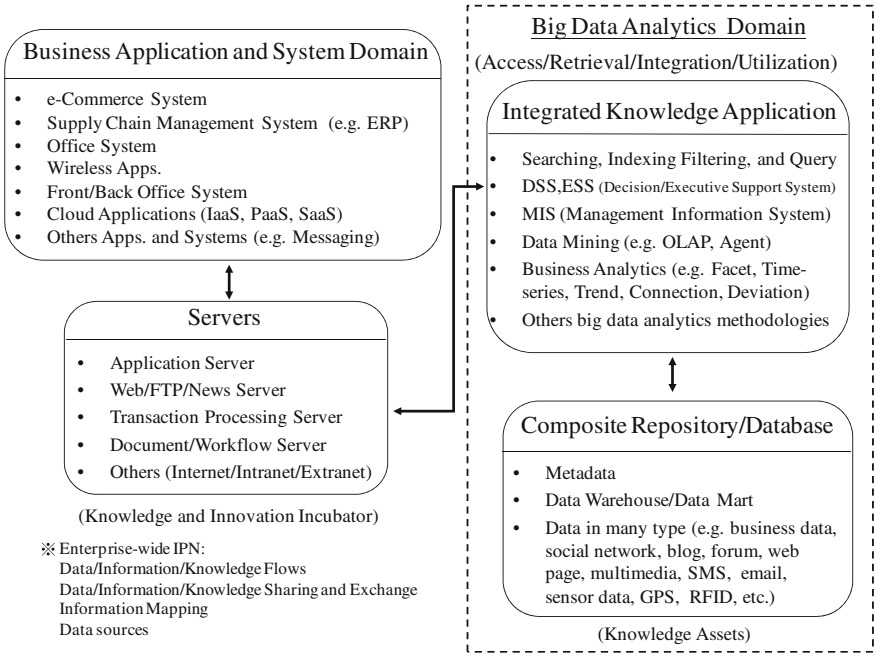


Fig. 9.3 Schematic system framework for big data applications

resources (applications). The role of servers can be a gateway or middleware that may reside in anywhere in the information processing network (public/private/hybrid cloud) that provide services.

To help enterprises organize information residing in multiple locations and deliver it to prospective users, directory service, indexing, and searching are required in integrated knowledge applications. In addition, the central or distributed data repository that provides or captures the data and information for employees and decision-making is very important in this application framework. In cloud application scheme, this would be the data center. The decision-making support system (DSS, ESS) is the driver that consolidates and directs the overall resources of the supply chain to the most mission-critical business activities. Therefore, the most important part to differentiate from traditional framework of supply chain application is the big data analytics domain. The information can be managed only when it is embodied as content, which represents a specific combination of information and a manageable data [13]. Given the Enterprise Resource Planning (ERP) system as an example, the ERP is good in managing processing data but fall short in providing intelligent tactics and strategic suggestions for decision-makings.

The big data analytics domain enhances the capabilities of mining, warehousing, extracting, and analyzing of heterogeneous data. For the analysis of data, there are growing numbers of analytics tools, algorithm, and artificial intelligence to improve a better extracting of heterogeneous data and turn into useful, valuable information

for supply chain decision-making and corporate knowledge assimilation. The proposed schematic structure for big data analytics combines automation, business rules, artificial intelligence, workflow, analytical tools and advanced messaging-analysis technologies to allow e-businesses to deliver information and to respond to customer requests rapidly and accurately [14]. The business applications will then couple with data analytics applications to conduct daily business operations, such as ERP, CRM, word-processing, spreadsheets, accounting, and so forth.

9.6 Conclusion

With the growing awareness of big data applications, enterprises have to recognize this competitive advantage and shift their focuses on building a robust knowledge-based system with big data analytics and applications capability. In such a way the enterprises are capable of quickly consolidating critical competencies and physical processes to gain competitive advantages easily.

On the other hand, the growing of strategic alliances and partnerships on a global scale that brought about the formation of inter-enterprise virtual organizations capable of leveraging the skills, resources, and innovative knowledge that re-side at different locations in a supply chain network. An organization's value chain consists of all activities performed to design, produce, market, deliver, and support its product and service. The value chain view of information processing network is introduced to enhance collaboration encourage innovation, boost productivity, achieve adaptivity, and increase the information system efficiency. The information process nodes are the keys to transfer data, information into values.

Big data, as the next frontier for innovation, competition, and productivity, in the era of big data, enterprises use their information processing networks and strive to become knowledge-enabled enterprises to ensure that all employees are able to utilize the knowledge and skills they need to meet their corporate goals. The proposed application framework stresses on the big data analytics domain to enhance the capabilities of mining, warehousing, extracting, and analyzing of heterogeneous data and turns into useful, valuable information for supply chain decision-making and corporate knowledge assimilation. It also combines automation, business rules, artificial intelligence, workflow, analytical tools and advanced analytics technologies to allow enterprises to deliver information and to respond to customer requests rapidly and accurately.

The proposed application framework, indeed, is integrated in terms of people focused on processes and values that ultimately respond to customer demand, but its success requires data that can integrate and support every exchange of information across the entire supply chain. Enterprises will realize that content is as important as technological framework to the enterprise application architecture. A well-designed and well-integrated knowledge-based SCM system with data analytics capability will improve existing supply chain performance and provide enterprise agility in the change business environment.

References

1. Computer Science Corporation (2000) "@ insights. A look at business transformation in today's e-world". Computer Science Corporation (CSC), 4–7
2. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute (MGI), San Francisco, CA, pp 1–137
3. Herodotou H, Lim H, Luo G, Borisov N, Dong L, Cetin FB, Babu S (2011) Starfish: a self-tuning system for big data analytics. In: Proceedings of the Fifth CIDR Conference
4. Porter M (1985) Competitive advantage: creating and sustaining superior performance. The Free Press, NY
5. Kuglin FA (1998) Customer-centered supply chain management: a link-by-link guide. AMACOM, a Division of American Management Association, New York
6. Gale T, Eldred J (1996) Getting results with the object-oriented enterprise model. SIGS Publications, Inc., New York
7. Curran TA, Ladd A, Keller G (2000) SAP R/3 business blueprint, understanding enterprise supply chain management. Prentice Hall, Inc, Sydney
8. Wang ET (2003) Effect of the fit between information processing requirements and capacity on organizational performance. Int J Inf Manage 23(3):239–247
9. Don Tapscott, (1999) Creating value in the new economy. Harvard Business Press, Cambridge, MA
10. LaValle Steve et al (2011) Big data, analytics and the path from insights to value. MIT Sloan Manag Rev 52(2):21–32
11. Senge PM (1990) The fifth discipline: the art and practice of the learning organization. Doubleday, Currency NY
12. Alen BJ (1999) Knowledge capitalism, business, work, and learning in the new economy. Oxford University Press, Oxford, pp. 56
13. Laugero G, Globe A (2002) Enterprise content services, connecting information and profitability. Addison Wesley, NY
14. Tiwana A (2000) The knowledge management toolkit, practical technique for building a knowledge management system. Prentice Hall, Upper Saddle River, NJ

Chapter 10
Website Navigation Recommendation Based on Reinforcement Learning Technique

Yin-Ling Tang, I-Hsien Ting and Shyue-Liang Wang

Abstract The explosive growth of the Internet has made information on the web large and complicated. If the structure of a website is not optimized, users could easily get lost and could not find the most important information at the first time. The adaptive website can present the information that users needed by analyzing the users' behavior. However, visitors may have different needs at different times. Most of recommended methods are not considerate of dynamic or time-dependent needs. This paper presents a recommender system based on reinforcement learning. We assume that five parameters are on recommendation, which include clicks of the page, time that spent on viewing the page, paths to find the page, hierarchy of the page, and the rank of the page. With the help of reinforcement learning to adjust the weight of five parameters, we aim to reduce the paths that user needed to find the object page.

Keywords Web usage mining · Adaptive web sites · Reinforcement learning

10.1 Introduction

In recent years, the internet has become the major source to access information. People have higher right to decide what information content to read, and want to access the desired information fast and efficiently. How to find the target information from the comprehensive data is a work to study. In contrast, websites need to consider the message presentation, helping people find the wanted information quickly. Therefore, websites need to be endowed with appropriate guidance for

Y.-L. Tang (✉) · I.-H. Ting · S.-L. Wang
Department of Information Management, National University of Kaohsiung Kaohsiung, 81148 Kaohsiung, Taiwan
e-mail: candy1989711@hotmail.com

S.-L. Wang
e-mail: slwang@nuk.edu.tw

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_10,

people while data growing rapidly. In order to have better guidance for users, we cannot ignore the structure of the website. But the optimal website architecture does not exist. If the website could adjust its presentation automatically to users' browsing feature to provide appropriate guidance, it could reduce the cost of searching, and keep the users satisfied. The website can be adjusted either by the user or by the owner of the network to access the wanted data more easily. But most users would skip the step of adjusting, and look for information directly [1]. Therefore, we need to use adaptive website to learn from users' browsing patterns, and adjust the website automatically to present the needed content [2].

Every website may not match users' need. The reason caused this situation may be that the design of the website was not appropriate, or users' need were changed [2]. The most common approach is providing a dynamic list to let the visitor link to popular page immediately. The website can find the appropriate recommendation page automatically by adaptive adjustment. But each of adjustments is an independent event, and all impacts of adjustment are by prediction.

Reinforcement learning explains the learning process of animals and children. In the learning process, learners will interact with the environment continuously to find the right active without supervision. It does not need a supervisor to monitor and teach in the process; the learning of the agent is only on the base of the mechanism of trial-and-error when interacting with the ambient environment. Each of actions impacts the subsequent states and actions. Through a series of learning, we can find out the most appropriate strategy for each state [3].

In this research, the academic website is our target. We apply optimization to adjust website for overall users' needs. Due to the lack of exact user information, we are using web mining to analyze log which is recorded by the server, and find out users' browsing patterns. Although optimization adjustment cannot reach every user's demand, it can improve overall performance of the website. The user who visits the website for the first time or unfamiliar with the site can find the information efficiently. In the research, we connect each adjustment together to learning how to adjust appropriately. Consequently, our research enables the website to learn as a human being with the aid of reinforcement learning, in order to make the optimal recommendation on the website to reduce the searching time.

This research has four purposes: (1) to analyze overall users' browsing habits and patterns. (2) to learn from the past adjustment experience by using reinforcement learning to offer recommendation according to users' need. (3) to reduce the paths of finding target information by recommendation. (4) to understand the importance of each browsing pattern of the website.

This paper is divided into five chapters: (1) to illustrate research background, motivation and purpose. (2) to discuss the definition of web mining, adaptive website, and reinforcement learning. (3) to describe the research system architecture and the process of data processing and to define the parameters in the reinforcement learning and to describe the method of adjustment and presentation. (4) to state the simulation operation situation of our system. (5) Conclusions.

10.2 Related Work

In this chapter, we introduce techniques that are used in this paper, including web usage mining, adaptive website, and reinforcement learning. Then we establish the method of our research by investigating past research.

10.2.1 Web Usage Mining

Web usage mining traces users' browsing behavior from the server automatically, and analyzes users' patterns provide with appropriate information. Users' browsing patterns can be utilized as the basis of website improvement. In doing so, website can present content more efficient and the interaction between the website and user become smoother.

Web usage mining will take out log files from the server, and preprocess the log files. Preprocessing includes data filtering, data cleaning, user definition, and session definition. If the data in the record did not meet our need, it would be deleted. To facilitate the following analysis, we preserve the needed fields of a data and remove the others. Next, we define the qualifications to distinguish difference visitors. And then we define the dividing point of sessions: 30 min is the usual interval [4]. After implementing above steps, the log is divided into different surf records to obtain users' browsing process, instead of a single record. Finally, we analyze the data that have been preprocessed to get the browsing patterns.

10.2.1.1 Data Source

Log is the major data resource of web usage mining. When users surf on the world wild web, the sever will save the browsing and accessing record in log. The default log format is NCSA ASCII Common Log File Format. It records user IP, host name, user name, date, time, access file, and file size. The format represent like [userIP][timestamp][method][url][httpversion][httpresult][size].

10.2.1.2 Mining Method

The important stage of web usage mining is the conduct of analysis. It finds out users' browsing patterns form numerous data and records. In pattern discovery step, the common techniques are Statistical Analysis, Association Rule, Classification, Clustering Analysis, and Sequence Pattern [5].

10.2.2 Adaptive Web Site

The more content displayed in website, the more complex pages it might links. Websites which provide different data and presentation by users' pattern can increase the surfing efficiency. It also lets the user find the information they needed quickly and promote the satisfaction for the website. An adaptive website learns the users' access patterns to improve the organization and presentation of website automatically [2].

There are two steps in the adaptive process. The first step is called behavior observation. This step finds users' browsing pattern by web usage mining, and model the patterns for following data matching [6]. The second step is called website adaption. It will set up many presented modules in advance, and then present the module that matched the browsing patterns. The second step is then re-divided into two types according to the target: content-level adaptation and link-level adaptation [7]. Content-level adaptation, also known as adaptive presentation, adjusts for website presentation. The purpose is to present different content for different kinds of users. Link-level adaptation is also called adaptive navigation support which adjusts for pages links in the website. It aims to provide navigation to prevent user lost in the website and find the information they need quickly.

Adaptive adjustment is divided in two ways according to adjusted method [2]. Customization adjusts pages immediately for each user's demand. Another method is optimization. Optimization would not adjust for individual user but for all users in order to make the website easier to use. Website adjust for over all users can improve new visitors viewing.

10.2.3 Reinforcement Learning

Reinforcement learning simulates the learning method of creature. It learns through failure and repeated practice to increase capability. Reinforcement learning is an unsupervised learning. In the learning process, it only needs to input data, and agent will learn to adjust automatically. Agent chooses the next action according to past experience, so it is suitable for dynamic environment. In an unknown environment, reinforcement learning will try different actions, and be interactive with environment. It is a learning method that uses trial-and-error and delay reward to find best strategy [3]. Figure 10.1 is the reinforcement learning module.

Reinforcement learning contains two roles. One is the environment that need to be changed. The other is the agent that is responsible for learning. In the learning process, after receiving environment state (s_t), agent will select an action (a_t) for the state and send to the environment. The current state (s_{t+1}) of environment will be changed by the action, and emerges the reward (r_{t+1}) for the effect. Next, agent will revises the value function for new state (s_{t+1}) and (r_{t+1}), and once again choose an action for the new state according to the value function. And then keep looping

Fig. 10.1 Reinforcement learning module

above steps. In the learning process, the effect of each action is the basis of conjoined selections. Through the learning experiences, the agent can find out the best action for different state, in order to reach the goal.

10.3 Methodology

This paper applies reinforcement learning to adjust recommendation for the website, helping users to find the target information quickly. Website will take optimized adjustment for overall users. It would integrate all users' browsing records for adjusting recommendation to reduce the searching time for users who might be unfamiliar with the website, and for new users. Besides page surf time and browsing time to consider, three supporting parameters are also taken into account in this paper, namely: path length to page, page level, and page current rank. The goal of our research is to reduce the paths to pages, so we add path to learning. Adding the latter two parameters is to consider previous browsing and adjustment in learning. In this study, the system is divided into three parts: extracting page pattern, learning recommendation weight, and adjusting recommendation list.

Figure 10.2 is the system structure of this research. When a user browse a website, apache records the user's accessing request automatically. Some steps are then proceed to analyze the user's browsing situation. First, we retrieve the log file from server, filtering unnecessary data in log, and then save the rest data into database. Next, we use data mining to identify the user's browsing patterns, and record them into the database too. After catching the users' patterns, system learns the value that needs to be adjusted in the recommendation by reinforcement learning. Each page will get a surf value by calculating these values. Finally, website presents the recommendation according to the surf value. The new log files that are recorded is going to be calculated again when the users start new browses. It would keep adjusting recommendation to fit most needs of user.

10.3.1 Extracting Page Patterns

In the experiment of our research, the data will be preprocessed first through steps as data retrieving, data splitting, data filtering, data cleaning, and data formatting,

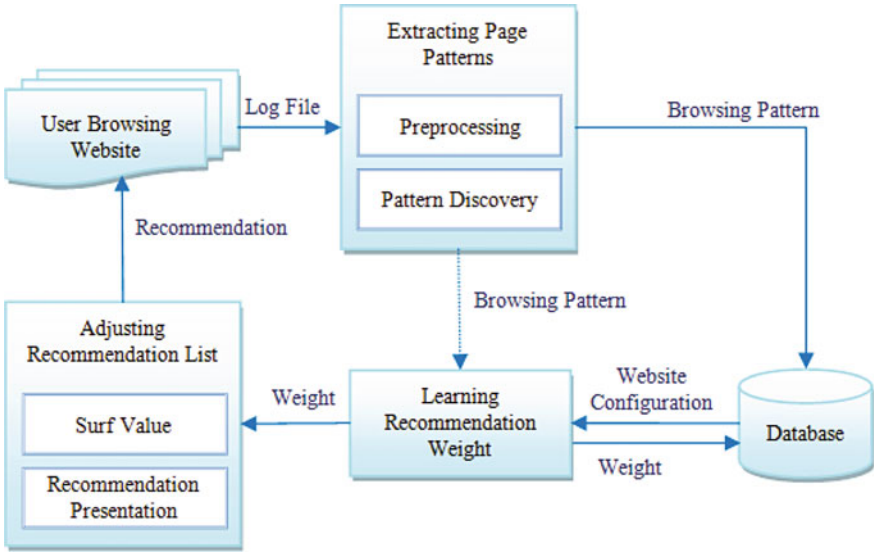


Fig. 10.2 Research system structure

and then go through user and session identification and store the data into database in the end. By running the steps above, we can get the users' patterns and the website structure. With knowing what function or information that the user needs, website can adjust itself with this information to improve the presentation.

(1) *Data Splitting*

It would generate redundant processes and increase running time for the compute if dealing with the oversized log file directly. So after retrieving the log file, we split the data into a suitable size. This paper is aimed to find out users' browsing trends. The shorter the interval of the analytic period, the better outcome it will result in. But the target website is not a huge one. It doesn't have sufficient data per day. Hence, we cut the file by three days.

(2) *Data Filtering*

The target website of this research experiment is encoded by PHP as a data filtering method to delete unnecessary data. To locate the path that the user finds the page, we preserve records which request ".php". Other records that request images or execution file, like *.jpg, *.png, and *.js, will be removed. We also filter out functional pages, and only keep the records that contain the presented page, index.php. Current PHP websites change pages by passing parameters, not linking pages directly. So we regard parameters as pages for the following analysis.

(3) *Data Cleaning*

After filtering the log file, we will remove the needless parts by data cleaning, and keep fields that will be used in subsequent steps. This paper focuses on analyzing browsing page. In order to find out paths which are in the same browsing page, we organized data on the basis of accessing date and accessing time before analysis. Therefore the fields that will be saved are user IP, timestamp, and url.

(4) *Data Formatting*

Data formatting turns data into the needed format to store the users' data into the database easily. We divide the timestamp into two fields: date and time. In url field, all records contain the same page, index.php. Hence we replace it with the passing parameters. And then we add "-" between each field to separate data. The format is [userIP]-[date]-[time]-[url].

(5) *User and Session Identification*

The common methods to identify users are using Cookie, user account, and user IP. In an announced website, user can search needed information without logging in. In the situation that lacks membership, the most common way to distinguish among users is using the IP address [8]. So we see the same IP as a user, and define 30 min as a cut-off point of Session. By using user IP and the interval time to manage log file, we integrate the browsing history together.

(6) *Page Parameter Retrieving*

We consider surf times and browsing time to know the importance of each page, and find out the object page that user wants. In this stage, we calculate clicks, stay time of each page. The stay time of a page is the period between the present request time and the next request one. The stay time of the final requested page is set as average time of all pages. The path length is the number of pages that are visited from first requested page to the last page. The path length of first request page is 0.

(7) *Page Database*

In the web mining step, the data will be stored into a database. The table recorded data that has been analyzed. Fields in table are browsing page, surf times, browsing time, average browsing time, path length to page, and average path length.

10.3.2 Learning Recommendation Weight

It is important to understand the interaction between the user and content in the website. We measure the website by tracing the pages that enter website and leave website, and also by tracing surf times, browsing time of each page. Each parameter can represent different features of the website, but unable to compare

with each other. Until now, we do not know which parameter is the most important one. Therefore, we let website learn automatically to find out appropriate weight of each parameter. Reinforcement learning cannot decide weight proportion at once, but decide it through a long-term learning by considering the past impacts of the weight adjustments. And then do following adjustment.

Besides tracing surf times and browsing time, we add other parameters such as path length to page, page current level, and page current rank into calculation. We suspect that recommend the page of longer path length will reduce more path length of whole pages. And the page at bottom level will be recommended prior to the page at high level. The surf times and browsing time in single does not represent that the importance of the page. So, previous rank will add into learning too.

In reinforcement learning module of this paper, the environment represents the browsing situation. First, the environment passes the data which was analyzed from the web mining to the agent. After getting current state of the page, the agent measures previous adjustment by evaluating the reward which was passed at the same time. Next, the agent calculates Q-value by SARSA algorithm and updates learning data. And then choose appropriate action for new state. The higher the Q-value is the higher property that action would be selected. After learning weights of each parameter, recommendation system sorts all pages in the website. Ultimately, system grabs top pages in the ranking to recommend, and presents them at a dynamic list.

SARSA algorithm formula:

$$Q(s,\ a) \leftarrow Q(s,\ a)\ + \alpha(R(s)\ + \gamma Q(s',\ a')\ -\ Q(s,\ a))$$

In this study, we simplify SARSA algorithm, setting α and γ as 1:

$$Q(s,\ a) \leftarrow R(s)\ +\ Q(s',\ a')$$

In the parameter weight learning state, the process of reinforcement learning is as followed. We will introduce the setting of sate, action, and reward in learning (Fig. 10.3).

(1) *Parameter Setting*

State: the current browsing situation of each page in website.

The website state that is traced by this study includes surf times, browsing time, and path length to page. In order to compare the data under the same standard, the three parameters will be standardized. Each data value represents the proportion of the page value in the whole website. For example, the surf time value is surf times of one page divided by surf times of all pages. The other two parameters are current level and ranking which will be standardized too for easier calculation.

State(Count, Time, Path, Level, Rank)

Action: agent chooses the parameter that need to be increased weight.

The actions that agent can select are adjusting weight. In a chosen action, the adjustments of the parameter are a fixed value. One parameter increase, the others' values will be reduced to keep the sum of all weights at a fixed value. The weight

1. Initial recommendation.
- In order to maintain the efficiency, we will calculate the recommendation for the first time.
2. Environment returns new state that influenced by recommendation.
- The state includes surf times, browsing time, path length, current level, and current ranking.
3. Environment returns the reward paired with the state-action.
- Reward is past path length minus current path length.
4. Update Q-value
5. Agent adjusts the weight of each parameter with Q-value.
- When the Q-values of actions are equal, pick one at random.
6. The recommendation system calculates ranking by weight which adjusted with action, and then recommend.
7. Back to the second step.

Fig. 10.3 Reinforcement learning recommendation process

of a parameter which has been selected will increase 4, others' weight will decrease 1 respectively then. And keep the sum of the weight at 100.

Action(WCount,WTime,WPath,WLevel,WRank)

Reward: the change of path length to target page.

The purpose of this study is to reduce the pages that are needed to go through when the user is searching information. Therefore, the change of the path length is the state-action evaluation standard. The calculation of the path length is limited in the target page. Other pages are not what we want to improve. Reward is calculated by past average path length minus current average path length. If the path length is negative, it means the adjustment of this time is not good, and vice versa.

Reward (TotalPathpast–TotalPathnow)

(2) *Parameter Weight Learning Algorithm*

In weight learning step, we input five parameters values into the learning, namely: surf times, browsing time, path length, current level, and current rank. And learn with algorithm as follows (Fig. 10.4).

In updating recommendation process, system firstly sums the path length of all pages to obtain a total path length of target pages to minus total past length to get the change of the path length. Regarding the change as the reward to evaluate the last adjustment, we input the reward into the reinforcement learning formula, and

Fig. 10.4 Weight learning algorithms

Reinforcement Learning Recommendation Algorithm

1. $\text{TotalPath}_{\text{now}} \leftarrow \sum \text{Path}$
2. $R \leftarrow \text{TotalPath}_{\text{past}} - \text{TotalPath}_{\text{now}}$
3. Update Q
4. $S \leftarrow$ Count, Time, Path, Level, Rank
5. Match S in RLmatrix
6. Select Max Q
7. Return A

update Q-value. Next, after regarding the surf times, browsing time, path length, current level, and current rank of page as the current state, we match the state in the reinforcement learning matrix. Finally, we select the action that has the highest Q-value, and return it.

After the state-action evaluation, the weight of each parameter will change. And we adjust the recommendation according to the new weight. Three days later, website gets new values of the five parameters, and run the algorithm again.

10.3.3 Adjusting Recommendation List

System traces users' browsing situation which contains the surf times, browsing time, and path length by carrying out web usage mining. And it grabs pages' allocation including current level and current rank from the database. By going through the reinforcement learning, we obtain the weight of five parameters which would affect the importance of pages. This study put the data that is found by mining and the weight that outcome by learning in the recommendation system. It can calculate the surf value of each page and rank the pages based on the surf value. Ultimately, it presents recommendation in the dynamic list in index page.

(1) *Surf Value*

To calculate the surf value needs five parameters and the weight of each one. The five parameters are surf times (Count), browsing time (Time), path length (Path), current level (Level), and current rank (Rank). Because these parameters had been standardized, they do not need to be processed additionally in the calculation. We multiply the parameters and its respective weight directly and sum up them. And then we obtain the surf value in that way. The formula is displayed as follows. Due to the value of the parameter and the weight did not execute other process, the weight represents the importance of each parameter.

Surf Value Formula:

$$\begin{aligned}\text{SurfValue} = {} & \text{Count} \times W_{\text{Count}} + \text{Time} \times W_{\text{Time}} + \text{Path} \times W_{\text{Path}} + \text{Level} \\ & \times W_{\text{Level}} + \text{Rank} \times W_{\text{Rank}}\end{aligned}$$

(2) *Recommendation Adjustment*

The website ranks all pages according to the surf value, selecting recommended page to present in the dynamic list. Recommendations are divided into two levels. The recommendations at the first level will be represented at home page. System filters out home page and pages at its next layer, and creates link to top three pages. As Fig. 10.5, the pages that can be selected to be recommendations are in the range of red dashed box. Next, the second level recommendations are presented at the next layer of home page. In order not to disrupt the original website architecture,

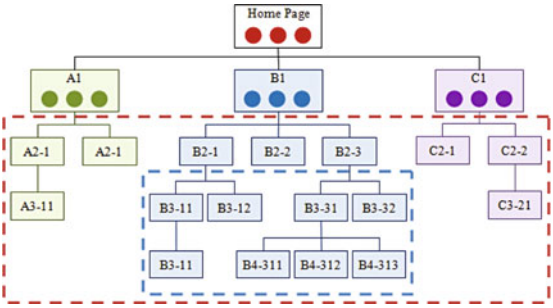


Fig. 10.5 Recommendation hierarchy schematic diagrams

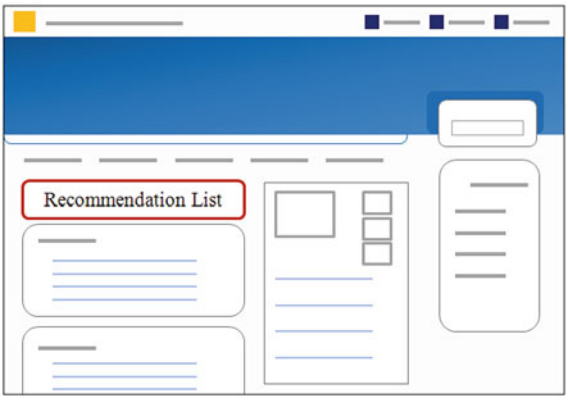


Fig. 10.6 Recommendation configurations at home page

we only choose pages that in the same category of the current page, and pick top three to recommend. In Fig. 10.5, we separate different category by color. Class B, for example, can only pick pages that in blue dashed box to recommend.

(3) *Recommendation Presentation*

In this experiment, the target website is a department website of a school. We improve the website based on the original site. Due to the users do not like changing screen [9], we do not disrupt the original website. The recommendations are presented in the dynamic list and displayed at an obvious place of the home page, see as Fig. 10.6. We aim to improve the browsing efficiency of whole website with a least change.

10.4 Experiments

The target website of this study is that of the Department of Information Management in National University of Kaohsiung. We use its log file as data resource. The website is an academic one whose primary purpose is to announce

information. The format of the log file is NCSA ASCII Common Log File Format. In this study, we retrieve the log file that recorded since May 2008 to September 2012. The size of original log file is 2.67 GB. We use log file to simulate system operation. The following is the simulated situation in January.

System Simulated Operation: Surf Times Increase

In early January, as the deadline for submitting proposal report was approaching, graduate students increased the surf times of "proposal report page". In this period, the surf times increased from 5 to 20 times. The average browsing time is still 5 min (300 s). The average path length is 5.5 pages. The number of pages is 120, and the surf time of whole website is 600 times. The average surf time is 5 times per page; and the average browsing time is 2.5 min (150 s). The average path length of target pages is 3.5 pages. Through standardization, environment returns the state as S (4, 2, 1.57, 0.5, 0.1).

The former weights of each parameter are 46, 42, 8, 2, 2. At this state, agent chose to increase the weight of surf times according to the Q-value. So, the weight of surf time increased 4, other weights decreased 1. After changing weights, the weights of parameters are order of 50, 41, 7, 1, 1. Through calculating by recommendation system, the surf value is 294.39.

$$\text{SurfValue} = 4 \times 50 + 2 \times 41 + 7.0.5 \times 7 + 0.5 \times 1 + 0.9 \times 1 = 294.39$$

Compare surf value of each parameter, "proposal report page" ranked the second. After filter home page and pages of next layer, it still at second. System got the top three pages to recommend directly. Hence "proposal report page" are in them. The new top three replace former recommendations and put at the dynamic list in the home page.

Three days later, the system learned again. The average path length of "proposal report page" reduced to 2.5 pages. The browsing time was still unchanged. Average path length of all target pages changed from 3.5 to 3.2. Average browsing time is 2.7 min. The website returned reward as 0.3 (Reward = 3.5–3.2 = 0.3).

Seeing that the path length is reduced, the reward is positive. It represents that the adjustment this time is appropriate. Next, agent re-calculated Q-value and selected next action according to the new state.

10.5 Conclusion

This paper proposes a method of page recommendation that is based on the reinforcement learning. Website recommends according to users' browsing patterns and website configuration. We consider multiple parameters into recommendation. The five parameters are surf times, browsing time, path length, current level, and current rank. After analyzing the log file to obtain user's pattern, the website adjusts weight of five parameters through reinforcement learning. And then to evaluate the adjustment automatically. Moreover, applying reinforcement

learning to interact with environment can consider temporality into recommendation. Finally to calculate each value of parameters and their corresponding weight. And then to recommend page for whole users to reduce the path length that needed when finding information. The recommendations of our research are adjusted with optimization. In the future, it can use the same method proposed by this paper to customize recommendation.

References

1. Schwarzkopf E (2001) An Adaptive web site for the UM2001 conference. In: Proceedings of the UM2001 workshop on machine learning for user modeling, p 77–86
2. Perkowitz M, Etzioni O (1997) Adaptive web sites: an AI challenge. In: Proceedings of international joint conferences on artificial intelligence, Nagoya
3. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge
4. Catledge L, Pitkow J (1995) Characterizing browsing behaviors on the World Wide Web. In: Proceedings of computer networks and ISDN systems, vol 27(6)
5. Srivastava J, Cooley R, Deshpande M, Tan P-T (2000) Web usage mining: discovery and applications of usage patterns from web data. In Proceedings of SIGKDD explorations, vol 1(2), pp 1–12
6. Perkowitz M, Etzioni O (1998) Adaptive web sites: automatically synthesizing web pages. In: Proceedings of association for the advancement of artificial intelligence, Madison, pp 727–732
7. Brusilovsky P (1996) Methods and techniques of adaptive hypermedia. User Model User-Adap Inter J 6:87–129
8. Pierrakos D, Paliouras G, Papatheodorou C, Spyropoulos CD (2003) Web usage mining as a tool for personalization: a survey. In Proceedings of user modeling and user-adapted interaction, vol 13, pp 311–372
9. Teíeni D, Feldman R (2001) Performance and satisfaction in aadaptive websites: an experiment on searches within a task-adapted website. J Assoc Inf Syst, 2(3):1–30

Chapter 11
An Approach for Hate Groups Detection in Facebook

I-Hsien Ting, Hsing-Miao Chi, Jyun-Sing Wu and Shyue-Liang Wang

Abstract In recent years, with the growth of social networking websites, users are very active in these platforms and a large amount of data is aggregated. Among those social networking websites, Facebook is the most popular one that has most users. However, in Facebook, the existence of Hate Groups is a very critical issue with the problem of abusing. Therefore, many researchers are devoting themselves to detecting the potential hate groups, using the techniques of social networks analysis and web mining. In this paper, we will propose an approach based on the techniques of social networks analysis and web mining to detect the potential hate groups. The data from Facebook are being processed. In the research, hate groups for 3C are selected as the training data. The social network structures and keywords of these groups will be treated as the features which will be used for discovering the potential hate groups in Facebook.

Keywords Facebook · Hate groups · Social networks analysis · Web mining

11.1 Introduction

With the rapid growth of the Internet communication techniques, the World Wide Web has become a very important platform for users to interact with each other. Through these platforms, users can easily share and spread information and ideas. In recent years, online social networking websites become very popular on which

I.-H. Ting (✉) · H.-M. Chi · J.-S. Wu · S.-L. Wang
Department of Information Management, National University of Kaohsiung,
Nan-Tzu District 81148 Kaohsiung, Taiwan, R.O.C
e-mail: iting@nuk.edu.tw

S.-L. Wang
e-mail: slwang@nuk.edu.tw

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_11,

users are able to share different information. YouTube (www.youtube.com) is a website for video sharing; Blogger (www.blogger.com) is designed for sharing the articles whereas Flickr (www.flickr.com) shares photos. Facebook (www.facebook.com), as well as MySpace (www.myspace.com) is a platform created for exchanging different kinds of messages, including photo, music and articles etc. These websites also become the busiest ones in the world.

Along with the people's interactions with each other on these social networking websites, more and more data and different information are aggregated. Among the information on these social networking websites, some are considered as negative ones because they could be utilized to attack and to slander. It is very interesting that this kind of information is always the information that spreads at the fastest speed.

Hate group is a group of users who attack, hate, and abuse particular objects, such as groups, companies, races or religions [1]. In the age of web 2.0, it is easier for hate group to affect users and to spread their ideas since the WWW is a public platform in which users can interact with each other. Especially in Facebook, the "Groups" function also provides a very good platform for those hate groups to share their ideas and to attract new users. Recently, the groups in Facebook have become a main target and therefore it becomes essential for us to develop a mean to locate hate groups.

In this paper, we intend to propose an architecture which is based on the classification technique of Data Mining. We will firstly observe the existing hate groups and the like groups in Facebook, and extract the features of these groups which include keywords that frequently used in them and the social network structure [6]. According to the extracted keywords and social network structure, the activity patterns can then be used to discover potential hate groups.

The paper is organized as follows: The first section is the introduction and the background of the paper. In Sect. 11.2, we will start to propose the approach of the paper which combines the techniques of Web Mining (text mining) and Social Networks Analysis. Due to the length limit of this paper, we are not going to discuss related literatures and works in detail. In Sect. 11.3, we will discuss more details about the approach and the experimental design to discover hate groups. The paper will be concluded in Sect. 11.4 along with the discussion of the future directions of this research.

11.2 The Approach for Potential Hate Groups Detection

In order to discover potential hate groups in the World Wide Web, some researchers have already proposed different means to deal with this problem. In 2003, Gerstenfeld et al. [2] analyzed 157 extremist sites and found links between most of these websites. In 2005, Zhou et al. [8] used software to assist the analysis of the content and links on the websites of hate groups. In their research, they found the main object of these websites is trying to spread and promote their ideas,

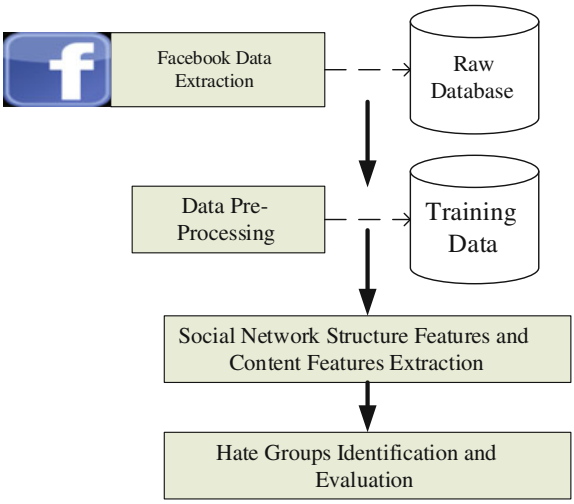


Fig. 11.1 The proposed approach for hates group detection

such as white supremacists and Neo-Nazis. Chau and Xu used the technique of social networks analysis and web mining to analyze the hate groups in the Internet, which is considered as the major breakthrough of the research in hate group detection. Wiil et al. [7] intended to analyze the hate groups according to the nodes in the social networks and to illustrate the social network structure. Warner and Hirschberg [5] are trying to detect the hate message by using the technique of text mining and semantic analysis. However, most of the researches concentrate on applying social networks analysis and link analysis for hate group detection. Therefore, in this research, we are going to propose a hybrid approach which combines the techniques of text mining and social networks analysis. Figure 11.1 is the proposed approach in this research based on the concept above.

In Fig. 11.1, we firstly will use the API that provided by Facebook to collect necessary data, such as all the messages in hate groups' wall. The collected data will be stored in the raw database. The second step of the approach is data pre-processing. The entire data in raw database will be processed to clean the data and extract useful data from the database. The third step of the approach is designed to extract social network structure features and content features. The last step of the approach includes is a serial of experiments to identify and evaluate hate groups The detail of the experiments will be discussed later in the paper.

11.2.1 Social Network Structure Features Extraction

Talking about the social network structure feature extraction, most of the common used SNA (Social Network Analysis) measurements will be extracted, including clustering coefficient, centrality, density and the average shortest path length [4].

These measurements are generated by using UCINet 6.0 (http://www.analytictech.com/ucinet/) which imports the preprocessed training data with the format of user Matrix.

11.2.2 Content Features Extraction

With regard to content features extraction, the message content in the hate groups will be pre-processing by using TF-IDF (term frequency-inverse document frequency) which is a very commonly used natural language processing technique. The keywords in the hate groups will be extracted and ranked according to the TF-IDF value for each keyword. Then, we will also extract keywords and measure the TF-IDF value for testing data. The similarity between the training data and testing data is measured by using cosine similarity [3].

According to the extracted social network structure features and keyword features, we will then composed the features as a group matrix of hybrid features. The matrix is shown below.

$$\text{Group} = \begin{bmatrix} \text{Density, Avg Shortest Distance, Clustering Coefficient,} \\ \text{Degree Centrality, Closeness Centrality, Betweenness Centrality,} \\ \text{Keyword Similarity, Is a late group} \left(\frac{\text{Yes}}{\text{No}} \right) \end{bmatrix}$$

The group matrix will be used for classification. The features will be treated as parameters in WEKA (http://www.cs.waikato.ac.nz/ml/weka/), which is a very famous and popular data mining tool. The experiment on measuring the performance of the proposed approach will be discussed in next section.

11.3 Experiment Design

In the previous section, we have introduced the group matrix which would be used for WEKA to classify groups for hate group detection. In order to evaluate the prediction performance, we design a serial of experiments with different number of keywords. Figure 11.2 shows the experiment design of this paper.

In Fig. 11.2, we intend to test the prediction performance of the hybrid features under different standards, namely: top 30 keywords, top 50 keywords and top 100 keywords. In WEKA, we selected two classification algorithms, which are J48 (C4.5) Decision Tree Classifier and Naïve Bayes Classifier. Furthermore, 10-flod cross validation is used for validating the accuracy of hate groups detection. After performing the WEKA, the evaluation results are shown in Table 11.1.

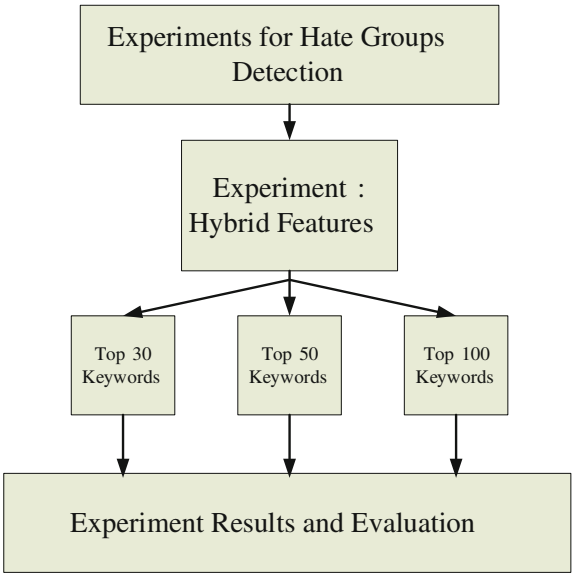


Fig. 11.2 The experiment design

Table 11.1 The performance evaluation summarization table of hate groups detection

Classifier	J48 (C 4.5)			Naïve Bayes		
Keywords	Top 30 (%)	Top 50 (%)	Top 100 (%)	Top 30 (%)	Top 50 (%)	Top 100 (%)
Precision	100	77.8	64.7	100	70	63.6
Recall	93.3	93.3	73.3	100	46.7	46.7
F-measure	96.67	83.33	66.67	100	63.33	60

Table 11.1 is the performance evaluation summarization table of hate groups detection. In this table, 100 % precision, recall and F-measure are shown as the best performance when using Naïve Bayes classifier with top 30 keywords. Furthermore, we can also found that the accuracy of performance is decreasing when the number of keywords applied is increasing. Therefore, the few keywords adopted for the hybrid features, the better accuracy of performance will be reached in the hate group detection.

11.4 Conclusion and Future Research Direction

In this paper, we have proposed an approach for hate groups detection in Facebook. The approach is a hybrid one which combines the features of social network structure and the features of keywords of message. In order to test the performance of hate groups detection, a serial of experiments have been performed to shows

acceptable results and performance. In the experiment, WEKA works as the classification tool as well as J48 and Naïve Bayes do.

In future research, we are going to test the proposed approach with only the features of social network structure or the features of keywords and compare the performances of hybrid approach. In addition, we can also select different classifiers, such as Neural Network etc., to test the performance, In the future, we also plan to develop approaches for detecting hate groups in different websites, such as Twitter, Plurk, and Blogger, etc. Furthermore, the researches of hate groups detection approach in the real world are also a very interesting direction.

References

1. Chau M, Xu J (2006) Mining communities and their relationships in blogs: a study of online hate groups. Int J Hum Comput Stud 65:57–70
2. Gerstenfeld PB, Grant DR, Chiang CP (2003) Hate online: a content analysis of extremist internet sites. Anal Soc Issues Public Policy 3:29–44
3. Salton G (1988) Automatic text processing: the transformation, analysis and retrieval of information by computer. Addison-Wesley Publisher, MA
4. Scott J (2000) Social networks analysis: a hand book. SAGE Publication, London
5. Warner W, Hirschberg J (2012). Detecting hate speech on the world wide web. In: Proceedings of the second workshop on language in social media, Montreal
6. Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge
7. Wiil UK, Gniadek J, Memin N (2010). Measuring link importance in terrorist networks. In: Proceedings of international conference on advances in social networks analysis and mining, Odense
8. Zhou Y, Reid E, Qin J, Chen H, Lai G (2005) US domestic extremist groups on the web: link and content analysis. IEEE Intell Syst 20:44–51

Chapter 12
Toward Crowdsourcing Data Mining

Hsin-Chang Yang and Chung-Hong Lee

Abstract Nowadays, crowdsourcing has emerged as a popular and important problem-solving approach. The major difference between crowdsourcing and traditional outsourcing lies on the people which tasks were outsourced. Those people involved in crowdsourcing are generally varied in knowledge, demographic properties, and number. Many applications and services have been developed to solve various types of tasks. However, these applications and services focus on providing platforms for outsourcing to the crowd. Little has been addressed so far on the management and usage of those information produced during the crowdsourcing process. Actually, as an emerging social network application and service, the data and social interactions created during crowdsourcing should carry important and valuable knowledge. This knowledge will develop various techniques for mining messages and information of crowdsourcing process. In this work, we address several approaches to discover useful knowledge from data created for and in crowdsourcing process. We hope the outcome of this research could help discovering usable knowledge from such emerging social network services and bring benefit in constructing crowdsourcing services.

Keywords Crowdsourcing · Text mining · Topic detection · Association discovery

H.-C. Yang (✉)
Department of Information Management, National University of Kaohsiung, Kaohsiung, Taiwan
e-mail: yanghc@nuk.edu.tw

C.-H. Lee
Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
e-mail: leechung@mail.ee.kuas.edu.tw

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_12,

12.1 Research Background

According to Nielsen's survey in 2012,[1] the growth rates of common social activity sites such as Facebook, Myspace, and LinkedIn dropped in 2012. In contrary, the most grown social services in last year include Pinterest, Blogger, Twitter, Tumblr, and Wikia, which provide content sharing and collaborative authoring. In this research, they are all referred as 'crowdsourcing' services, whose definition is somewhat restricted in past decade. Howe [3] defined 'crowdsourcing' as '…represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call'. This definition just widens the traditional outsourcing process to incorporate online environments and processes. On the other hand, Brabham [1] adopted a broader definition as which 'crowdsourcing is an online, distributed problem-solving and production model'. In this regard, it fits our previous definition of crowdsourcing, in which all services incorporating collaborative content creation and task completion are included.

Most of contemporary crowdsourcing services provide platforms for task announcement and worker recruitment. A famous example is Amazon's Mechanical Turk.[2] Besides such task-worker matchmaking platform, other services are also possible by crowdsourcing, such as crowdvoting and crowdfunding. Message exchange is necessary for crowdsourcing. Various types of messages are available throughout the crowdsourcing process, e.g. usage pattern, user profile, link data, tags, and text messages. Manipulating and mining of such messages are seldom discussed in past. Therefore, it is unclear about the plausibility and effectiveness of data mining, especially text mining, techniques on such data produced during crowdsourcing processes.

12.2 Research Goals

In this research, we will try to achieve the following goals regarding crowdsourcing data mining:

1. To establish schemes for crowdsourcing data management and visualization.
2. To develop kernel techniques for crowdsourcing data mining, such as topic detection and relation discovery.
3. To establish a platform to demonstrate the effectiveness of proposed methods.

We expect that our research will provide a uniform scheme to levitate the data usage in crowdsourcing.

[1] http://blog.nielsen.com/nielsenwire/social/2012/

[2] http://www.mturk.com

12.3 Research Methods

We will describe the major steps of this research in the following:

Date collection and processing The volume of data, in various types, produced in crowdsourcing process is usually large. We only focus on textual data in this research. Two types of textual data will be collected, namely messages and profiles. Profiles are used to provide demographic and social attributes of messages which are the major sources of mining process. We will develop several approaches to clean, reduce, and normalize these messages, as well as attaching attributes.

Data clustering and classification We will apply self-organizing map (SOM) algorithm to cluster messages to discover the relations among messages. Various SOM implementation, such as classical SOM [4], growing hierarchical SOM [2], and topic-oriented SOM [5], will be used to verify their effectiveness. We also perform clustering process with profile data to obtain demographic clustering of messages.

Topic detection For further investigation of relationships among messages, the topics of messages will be discovered through a topic detection process. Here a topic is a set of keywords that could possibly describe the main idea of a message. We will develop a detection scheme based on message clustering result to discover semantic terms. These topical terms will then be used to perform thematic categorization of both messages and profiles.

Association discovery The purposes of this process is to discover the relations among messages, users, and topics. Since the clustering process should be able to discover relations among messages, users, and topics, respectively, the goals of this step is to find the association across messages, users, and topics.

Application platform implementation In the final stage of this research, we will implement a platform to demonstrate the usage and applicability of our proposed crowdsourcing mining process. We plan to establish a disaster information coordination platform, which incorporates real-time reports from users. Trends, associations, events, and other useful knowledge regarding disasters could be discovered and disseminated using this platform.

12.4 Expected Result

We expect to achieve the following results in this research:

1. Gather and process crowdsourcing data from various platforms for further researches.
2. Complete development of topic detection and association discovery algorithms, as well as other derived algorithms, such as event detection, automatic summarization, spam detection, and content recommendation, etc.
3. Establish a experimental platform for disaster information coordination.

12.5 Conclusion

Crowdsourcing is a new way for problem solving in Web era. However, data management and usage are seldom discussed in such process, let alone knowledge discovery from such data. In this work, we address a proposal to establish a framework for mining crowdsourcing data mainly based on text mining techniques. Several techniques for mining crowdsourcing data will be developed. We expect the result of this research could be beneficial for applications and researches on crowdsourcing and broaden its usage.

References

1. Brabham D (2008) Crowdsourcing as a model for problem solving: an introduction and cases. Convergence: Int J Res New Media Technol 14(1): 75–90
2. Dittenbach M, Merkl D, Rauber A (2000) Using growing hierarchical self-organizing maps for document classification. In: Proceedings of the 8th European symposium on artificial neural networks (ESANN'2000), Bruges, 7–12
3. Howe J (2006) Crowdsourcing: a definition. http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html
4. Kohonen T (2001) Self-organizing maps. Springer, Berlin
5. Yang HC, Lee CH, Ke KL (2010) TOSOM: A topic-oriented self-organizing map for text organization. World Acad Sci Eng Technol 41:1100–1104

Chapter 13
Wireless Security Analysis Using WarDrive Investigation in Kaohsiung Areas

Hanwei Hsiao, Tienhe Chang and ChihChe Chang

Abstract With the rapid developments in wireless LAN (WLAN), all types of mobile internet applications and the number of its users get the fast growing. This phenomenon is more obvious in metropolitan areas, and the respondent wireless internet security problems are getting worse. This study would like to look into the WLAN security problems in Kaohsiung metropolitan areas using the wardriving technique, and collects the detailed data of WLAN Access Points (WLAN APs), especially including the items of encryption or not, the types of encryption, the Service Set Identifier (SSIDs), and the Received Signal Strength Indicator (RSSI) to show the correlated wireless LAN security issues. These data will be compared to the related ones of others countries for offering the objective comparisons in WLAN security. In this study, we also randomly grabbed 227 wireless WPA/WPA2 encryption packets and analyzed its passwords strength by GPU technique, and the 56.24 % are the simple passwords and easily cracked by malicious behaviors which show that another worthy issue of the WLAN in Kaohsiung metropolitan areas. This study show the data and issues of WLAN security empirically which offers relatively security statistical data for further research and promotes the secure awareness of wireless internet users.

Keywords Wireless security · Wardrive · Passwords · Encryption · Wireless network

H. Hsiao (✉) · T. Chang · C. Chang
Department of Information Management, National University of Kaohsiung, Kaohsiung, Taiwan
e-mail: hanwei@nuk.edu.tw

T. Chang
e-mail: l1003303@mail.nuk.edu.tw

C. Chang
e-mail: m1013311@mail.nuk.edu.tw

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_13,

13.1 Introduction

According to IDC [1] latest data, in the worldwide smartphone market, vendors shipped 216.2 million units in 1Q13, which marked the first time more than half (51.6 %) the total phone shipments in a quarter were smartphones, it is because WLANs offer various advantages over conventional wired networks, viz., efficiency, service, handiness, and cost [2]. By the data from 2001 to 2013 of wigle.net, there are 94,628,906 WLAN APs worldwide until 28th April 2013 [3].

People are apt to ignore that the WLAN signals can be easily detected while enjoying the convenience and the present WLAN signals are now highly available because of the radio wave transmitted at random. By the street test from the crossroads A and B, there are about 155 WLAN APs in 300 m × 300 m area which means 1722 APs/km^2, and there are still school and park in this area. It shows that low threshold and easiness of WLAN APs detection because all types of detection applications develop so well, and smart mobile devices can be with multiple WLAN detection application programs.

To replace the insecure WEP encryption, WPA and WPA2 was respectively formulated by IEEE in 2003 and 2004, however, WLAN users often set the simple passwords for convenience which called the lazy passwords, and that were easily cracked by its less strength. With good excessive radio wave, the malicious attacker could log in victims' WLAN by just standing in the street, and this would led to the huge connection fee, the theft of important data… without any awareness, so the excessive radio wave is the primary issue of WLAN security.

This study collects and analyzes the SSIDs, encryption or not, the RSSI, the encryption types of WLAN APs by wardriving technique. Among those data, encryption or not is the most important protection measure of WLAN, the degree of excessive radio wave is the first key factor for attackers, and types of encryption determine the difficulty for cracking. With this, we can collect practical data of WLAN APs in order to test the kinds of security issues, examine the reliability and security of WLAN, and analyze the coverage of WLAN.

13.2 Related Works

13.2.1 The Security of WLAN

The growth of WLAN with all kinds of mobile devices, the security issue is getting worse as previous mentioned such as the features of invisibility, highly detectable, and availability of invasion. If the wireless network has no adequate protection system, any computer within range may access the network [4]. So, security is a serious concern because the wireless medium is open for public access within a certain range [5]. The wireless networks are more vulnerable than wired networks because the data is transmitted through the broadcast radio technology [6]. From

these points, facing with the excessive waves of WLAN APs is the top security issue while people enjoy the convenience of wireless communication. Security risks in wireless environments include risks of wired networks plus the new risks as a result of mobility [7]. We can see that the present wired networks got lots of problems to be dealt with, and the WLAN makes the security issues worse because of the radio wave goes in all directions. Since if not secure, this technology will provide an remotely accessible attack surface distributed throughout many homes and businesses [8]. Francisco points out that computer networks and more specifically wireless communication networks are increasingly becoming susceptible to more sophisticated and untraceable attacks [9]. We can test the Received Signal Strength Indicator (RSSI) in a simple way: −30 dbm (very good) to −127 dbm (poor), it means that the RSSI value will be −30 dbm if a WLAN AP is next to the researcher's mobile device, on the contrast, the RSSI value will be −127 dbm if a WLAN AP is 200–300 m away. The values above are relative, they differ in various antennas and chips sensitivities which means that the same mobile device shows different dbm values with various antennas, and it would be an important concern for those consumers who make the set-ting in improper ways.

13.2.2 WLAN Encryption and Passwords Analysis

The encryption between WLAN APs and mobile devices must be done because Lashkari points out that encryption is optional in 802.11 WLANs, but without it, any other standard wireless device, can read all traffic in network [10]. The most important form of protection lies in encryption methods for wireless networks and research results indicate a disappointingly high proportion of wireless networks using inadequate protection [4]. Based on the WLAN communication mechanism by time, there are three encryption protocols: (WEP) Wired Equivalent Privacy-WEP: IEEE, 1999), WPA (Wi-Fi Protected Access: IEEE, 2003), WPA2 (Wi-Fi Protected Access Version 2: IEEE, 2004) [10]. The Wi-Fi Alliance defined the protocol of WPA/WPA2 in response to several weaknesses researchers had found in the previous system: Wired Equivalent Privacy (WEP) [11]. The IEEE announced the standard of 802.11i which is the more secure WPA2 encryption protocol in 2004. By the street test, it shows that the main WLAN APs encryption is WPA/WPA2 nowadays. However, as discovered vulnerabilities, the security of WPA/WPA2 is threatened [12]. In the analysis of WPA/WPA2 passwords, we have to take the WLAN encryption packets, and the only way is deauthentication which disconnects the communication between WLAN AP and its clients. Now we deauthenticated a client from the wireless network, then the client would re-exchange the WPA key [13]. This is the timing we grab the WLAN encryption authentication packets in milliseconds.

Owing to the low speed of cpu-based method, this study adopts the highly-paralleled computing GPU [12]. Besides, there is one method closely related to the SSIDs which called Time-Memory Trade-Off (TMTO) in cryptanalysis.

The TMTO attack reduces the time of cryptanalysis by using data stored in memory [14]. This pre-calculating means that we just do the comparison without mass computing in cryptanalysis.

13.2.3 Wardriving and Related Surveys

Wardriving is the act of searching for Wi-Fi wireless networks by a person in a moving vehicle. Wardriving was first developed and used in this manner by Pete Shipley in April 2001 [15]. The study using wardriving must keep away from the conditions: theft of services, denials-of-service, and theft of information [15]. Now, there are some foreign researchers implement wardriving activities, such as Macao- ISACA [16], Norway [17], New Zealand [18], Hong Kong-HKWDC [19], Croatia [4], and the wigle.net [3] in order to promote the awareness of WLAN security. The Table 13.1 is the wardriving data of each area.

From the data above, there are still many users setting the APs open to the public or taking the WEP method, and it provides the chances for malicious attackers because they can find and hack APs in the remote for illegal behaviors.

13.3 Design

We collect and analyze the data of WLAN APs using wardriving technique by mobile devices to establish the SSID database and define the usual default SSIDs by the number of its appearances. Then, we grab the WLAN encryption packets which contain the data of WPA/WPA2 using mobile devices and external antennas, and analyze the passwords to verify the insecurity of WLAN. The design of this study is shown as in Fig. 13.1.

Table 13.1 The wardriving of 6 areas

Areas	Unencrypted (%)	WEP (%)	WPA/WPA2 (%)
Macao 2010	13.2	27.5	59.3
New Zealand 2011	16.8	8.2	76.0
Hong Kong 2012	16.6	20.9	62.5
Norway 2012	14.2	21.6	59.5
Croatia 2012	5.00	37.0	59.0
Wigle.net global 2001–2013/4/28	17.8	22.9	44.2

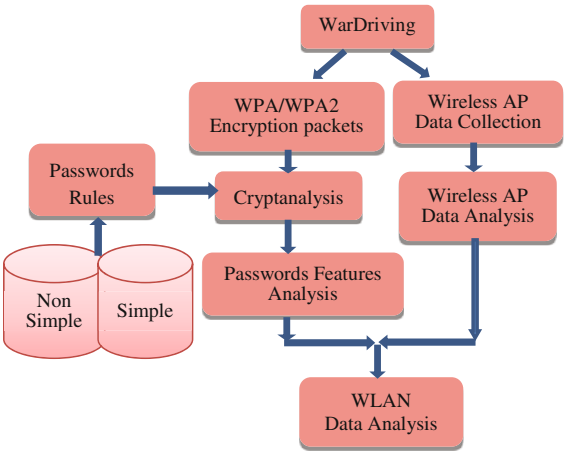


Fig. 13.1 The design diagram of the study

Table 13.2 The hardware list of wardriving

Mobile device	External chip	External antenna	Vehicle
NoteBook	RL3070	12dbi	Car
Samsung S2	Built in antenna	Built in antenna	Car, motorbike

13.3.1 Data Collection and Analysis of WLAN- WarDriving

This study makes use of the feature of radio wave speed close to the light, combining wardriving technique by mobile device in the vehicle to collect the data of WLAN APs in certain areas. Within the range of the antenna, we can collect the related data of WLAN APs. The wardriving hardware was listed in the Table 13.2.

13.3.1.1 Routes of Wardriving

The wardriving regions are divided into two major parts which are the north and the south part of Kaohsiung metropolitan areas based on the Kaohsiung train station railroads. The two major divisions are: the north part is the newly rise business areas, the south part is the old metropolitan areas, and then we collect the data from these two major parts in the intersectional driving courses. The south routes of empirical data collection of WLAN APs is shown as Fig. 13.2.

13.3.1.2 The Power of Antennas

Another key point is the power of antenna; it is an important factor for the volume of data. With the advance in technology, the power of antenna of WLAN APs with

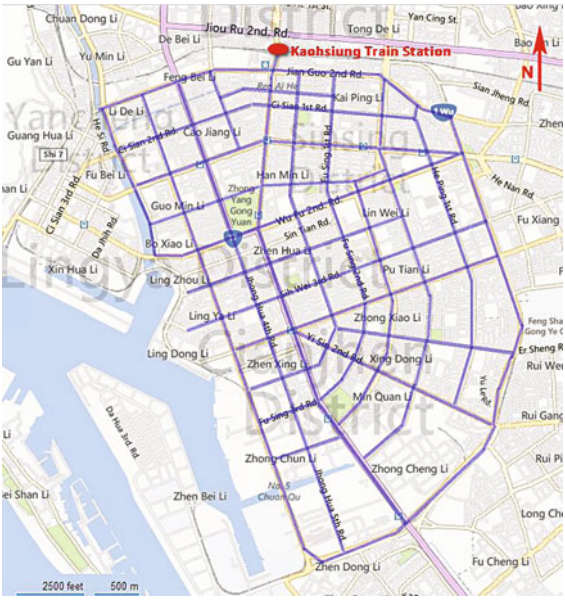


Fig. 13.2 The south routes of empirical data collection of WLAN Aps

multiple functions has the great help in data collecting which means the more power of the antennas, the more data volume. The more power of antenna means the higher RSSI. Empirically, we received the WLAN APs with the SSID of nukedutw from 800 meters away, and this is the best example of the more power, the higher RSSI. The street test data could be shown in the Google Earth, but its location is not accurate even though with the help of GPS and the GPS gets errors itself, so the location is just for reference, not an exact value.

13.3.2 The Analysis of WLAN APs Passwords

This study not only collects and analyzes the data of WLAN APs in the street using wardriving technique but also grabs 227 WLAN APs encryption authentication packets which are primary the WPA/WPA2 methods for cryptanalysis in order to investigate the security of WLAN passwords. For showing the insecurity of WLAN, we analyze the passwords adopting the method of brute force attack using GPU [12], and most of them are simple passwords which are what we called lazy passwords.

13.4 The Empirical Analysis of WLAN Security

There are 27847 unique WLAN APs data after the filtering, and we list the items that would be analyzed for this study as shown in Table 13.3.

Table 13.3 The analysis items of wardriving in Kaohsiung areas

Mac address	The basis of filtering	Kick off the duplicated
SSID	Analytic factor of WLAN encryption	Define the customized SSID
RSSI	Necessary factor for malicious behaviors	Average −84.54 dbm (good)
Encryption Types	WEP——————————9.1 % WPA/WPA2——————————75.8 %	
Encryption or not	Unencrypted: Top 1 choice of attacks	15.1 %

From Table 13.3, we can see the items to be analyzed, Mac Address is the basis for excluding repeated data, and the SSID is the name of AP, except that, the SSID is an important analytic factor of WLAN encryption authentification because the SSID is encrypted with the passwords, so analyzing passwords equals analyzing the SSIDs. The average RSSI of −84.54 dbm in the data of 27847 APs is a very excellent value, and this is also good to the malicious attackers. With radio wave, that is what we can contact the WLAN APs or mobile devices from the air, and execute the kinds of malicious behaviors. We often received the good WLAN wave of −60 to 70 dbm, and the best value of −30 to 40 dbm indoors from our street test. According to the test of 5 years old AP, we still can receive an excellent value of −30 to 40 dbm from 170 m away, and the present APs are more advanced which show that the very insecure part of the whole WLAN environment.

From Table 13.4 and Fig. 13.3 [3, 4, 16–19], it show that the statistics of WLAN APs in the seven areas, and we can see that WPA/WPA2 encryption methods are highly used to protect WLAN communication. There are at least about 60 % usages of WPA/WPA2 in the seven areas except the data of wigle.net which means that people has the basic cognition of WLAN security. As for the WEP, except the areas of Kaohsiung and New Zealand, the rest areas are about 20–37 % which is proved to be insecure, that's why WPA/WPA2 established, and the value of WEP in New Zealand is about 8.2 %, the lowest among these areas, and the 76 % of WPA/WPA2 is the highest among these areas which means people in New Zealand area is relative secure in WLAN. For the part of open APs, it is about 5–16 % to be open to the public, and provides the chance for the

Table 13.4 The wardriving data analysis in seven areas

Areas	Unencrypted (%)	WEP (%)	WPA/WPA2 (%)
Macao 2010	13.2	27.5	59.3
New Zealand 2011	16.8	8.2	76.0
Hong Kong 2012	16.6	20.9	62.5
Norway 2012	14.2	21.6	59.5
Croatia 2012	5.00	37.0	59.0
Kaohsiung 2013	15.1	9.1	75.8
Wigle.net global 2001–2013	17.8	22.9	44.2

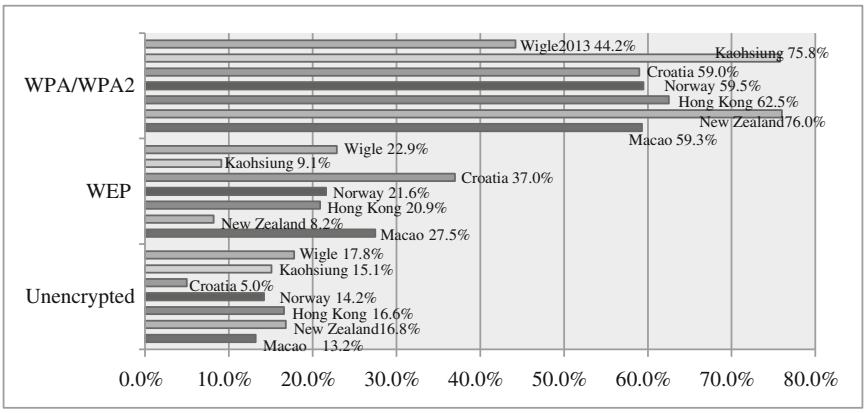


Fig. 13.3 The wardriving data analysis in seven areas

Table 13.5 The customized SSID of wardriving

Rank	SSID	Num
1	Dlink	956
2	P874	587
3	EDIMAX	493
4	Default	422
5	AndroidAP	391
6	APTG Wi-Fi	371
7	ASUS	362
8	CHT Wi-Fi auto	359
9	DSL-6641 K	282
10	HTC portable Hotspot	246

attempted. It will be quite a big volume if this proportion of 15.1 % open APs of certain areas in Kaohsiung enlarged to the whole City which is a challenging number of WLAN security.

We can find that the top 10 customized SSIDs are all default values from the 27847 APs data. As shown in Table 13.5 and Fig. 13.4, it shows the great secure worry that the top one default SSID is dlink with the number of 956 because this kind of encryption of WPA/WPA2 contains the default SSID and lazy passwords which is the main target of TMTO attack. This attack is based on the pre-computing in cryptanalysis, the malicious attacker just compare the results to the encryption packets with the speed of millions/s in cryptanalysis. Now, there are organizations collect kinds of the default or customized SSIDs combining the lazy passwords to build the database which will be a great threat for the people with lower WLAN security awareness if the data were utilized by malicious attackers.

Except the empirical investigation of WLAN APs, this study also grabs the WPA/WPA2 encryption packets to analyze its security strength. The present WPA/WPA2 encryptions are highly complicated algorithms, and it could protect

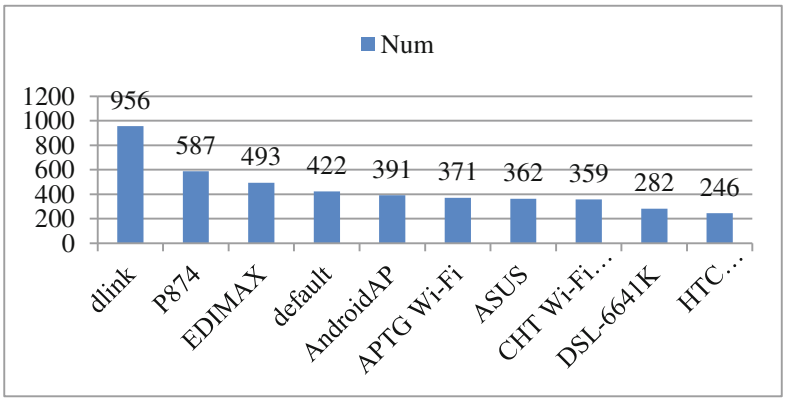


Fig. 13.4 Top 10 customized SSIDs of wardriving

the WLAN well theoretically. These simple passwords result in the WLAN security vulnerability because the users take the simple passwords for convenience and easy to remember. This study collected 227 WLAN encryption packets at random geographically, and analyzed with four major categories of passwords combinations-local phone number, mobile phone number, dictionary files, and 8 digits number. By the hardware of this study, it just took few seconds or minutes to crack these passwords, and its cracking percentages can be reach to 56.24 %. From this, we can see there are many potential WLAN security concerns, although the users in Kaohsiung metropolitan areas use the more secure WPA/WPA2 encryption methods (75.8 %).

13.5 Conclusions and Future Research Directions

From the wardriving, we get the information that every aspects of values are all against to the internet users, and the most users rarely do the setting of the WLAN APs because of convenience. And this default values are public information. This default SSID values are the 1st target of TMTO attack [12] in WLAN crypt-analysis. Furthermore, we got the average RSSI value of −84.5 dbm, and this is a good signal index, and we can reach our detective range even to 800 m. The consumers always choose the WLAN APs of multiple function with high power antennas just for the ease of usage and connection strength. Putting these conditions together, strong receiving ability for both WLAN APs and its clients, it provides a "excellent friendly" condition for malicious attackers, and they can do the malicious behaviors in a more distant, secret, and flexible spot.

People have to think of the security issues relative to its convenience while enjoying the WLAN, and the percentage of WLAN usage and universality will keep going up. The most applicable protection in present WLAN security is

Table 13.6 The recommendations of WLAN

Items	Recommendation	Prevention
SSID	Define it yourself	TMTO attack
RSSI	Bigger is not better	Easy to be detected
Passwords	Don't use simple one	Easy to crack
WEP encryption	Don't adopt It	Easy to be attacked
OPEN APs	Adopt WPA/WPA2	Easy to crack

WPA/WPA2 encryption, but it caused the new security crisis because of inadequate usage.

As previous noted that the best protection of WLAN is passwords which could protect the security of WLAN well. It can be easily cracked, and the reason is the "human factor" which is for convenience. And these passwords can be guessed, that's why we suggest set the more complicated passwords.

In sum, according to this study, we conclude that there are 5 major elements of WLAN security which are the SSIDs, the RSSI, passwords setting, old WEP encryption, and open APs, we suggest the WLAN users could check them individually as shown in Table 13.6 for the purpose of upgrading the security awareness of WLAN.

The progress of technology is beyond imagination, especially the wireless technology. The transmitting distance of WLAN APs is getting more distant and its applications are even more prosperous and diverse everywhere. This study takes those present high technology to investigate the WLAN security in Kaohsiung metropolitan areas. It shows us the convenience and insecurity of technology, but there are still some restrictions in this study for further research which contains stronger Antenna, higher parallel computing GPU, higher density of wardriving investigation, and mobile phone with external antenna.

References

1. Llamas R, Restivo K, Shirer M (2013) More smartphones were shipped in Q1 2013 than feature phones: An industry first according to IDC. Accessed http://www.idc.com/getdoc.jsp?containerId=prUS24085413
2. Bhatia V, Gupta D, Sinha HP (2012) Analysis of dictionary attack on wireless LAN for different nodes. J Inf Syst Commun 3(1):167–169
3. "Wigle.net," *General Stats* (2013). Accessed https://wigle.net/gps/gps/main/stats/
4. Janić D, Perakovic D, Remenar V (2012) An analysis of wireless network security in the city of Zagreb and the Zagreb and Karlovac counties. In Proceedings of the 7th international scientific conference on ports and waterways-POWA 2012, 2012
5. He C, Mitchell JC (2004) Analysis of the 802. 11i 4-way handshake. In: Proceedings of the 2004 ACM workshop on wireless security wise 04, pp 43–50
6. Luminita D (2012) Wireless LAN security-WPA2-PSK case study. In: 2nd world conference on information technology (WCIT-2011), vol 1, pp. 62–67
7. Gupta A, Kumkar V, Shrawne S, Tiwari A, Tiwari P (2012) Vulnerabilities of wireless security protocols (WEP and WPA2). Int J Adv Res Comput Eng Technol 1(2):34–38

8. Goodspeed T, Bratus S, Melgares R, Smith SW, Speers R (2012) Api-do: tools for exploring the wireless attack surface in smart meters. In: The 45th Hawaii international conference on system sciences, pp 2133–2140
9. Aparicio-Navarro FJ, Kyriakopoulos KG, Parish DJ (2011) An on-line wireless attack detection system using multi-layer data fusion. In: Proceedings of 2011 IEEE international workshop on measurements and networking (M&N), 2011
10. Danesh MMS, Lashkari AH, Samadi B (2009) A survey on wireless security protocols (WEP, WPA and WPA2/802.11 i). In: The 2nd IEEE international conference on computer science and information technology, pp 48–52
11. Ahmed S, Belali MH, Mahmud I, Rahman S, Sakib N (2012) WPA 2 (Wi-Fi protected access 2) security enhancement: analysis and improvement. Glob J Comput Sci Technol 12(6):83–89
12. Jin Z, Liu Y, Wang Y (2010) Survey on security scheme and attacking methods of WPA/WPA2. In: Proceedings of the 6th international conference on wireless communications networking and mobile computing (WiCOM), 2010
13. Ambavkar PS, Patil PU, Meshram BB, Swamy PK (2012) WPA exploitation in the world of wireless network. Int J Adv Res Comput Eng Technol 1(4):609–618
14. Oechslin P (2003) Making a faster cryptanalytic time-memory. In: Proceedings of the 23rd annual international cryptology conference, 2003
15. Muhammed Azhar YM (2010) Wardriving. In: Seminar report at computer science engineering of Cochin University of Science and Technology, 2010
16. Manetic M, Isaca M, Pisa HK (2010) Wardriving Macao 2010
17. Svendsen G (2012) Security state of 802. 11 wireless networks a study of WLANs in five Norwegian cities. Universitas Bergensis
18. Nisbet A (2012) A table of four cities wireless security growth in New Zealand. In: 2012 international conference on computing, networking and communications (ICNC), 2012, pp 1167–1171
19. Fong K (2012) Capital weekly. Accessed http://kencept.blogspot.tw/2012/06/wi-fi.html

Chapter 14
Guanxi Buying in the Social Media Environment

Cathy S. Lin and Shin Yan Lu

Abstract In the online marketspace, social media has played an indispensably role in changing the way of online shopping and auctioning. Previous studies have found the interpersonal relationship is the key factor to a success social media business. This study aims at exploring group buying in social commerce environment from the Chinese "guanxi" perspective. Guanxi is a kind of interpersonal interaction that individuals tend to interact through the *ren-quin* (favor), and individuals have to give and save *face* for their friends to show their social friendship. Chinese culture is differentiated from Western culture in many aspects. Therefore exploring the impact of Chinese guanxi factors in affecting consumer participating in the group buying decision making has its research originality to bring the Chinese interpersonal relationships Guanxi into social media context.

Keywords Social media · Guanxi · Ren-quin · Face · Group buying

14.1 Introduction

According to Market Intelligence Center (MIC, 2013) survey, the output value of Taiwan's e-commerce exceeded NT$6,600 million and the annual growth rate of 17.4 % in 2012, estimated to break 1 trillion in 2015. The type of online shopping

C. S. Lin (✉) · S. Y. Lu
National University of Kaohsiung, 700, Kaohsiung University Rd, Nanzih District, Kaohsiung 811 Taiwan, Republic of China
e-mail: cathy@nuk.edu.tw

S. Y. Lu
e-mail: blue66656@hotmail.com

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management,* Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_14,

are diversified, in recent years, it is worth noting that Taiwan's group buying significantly improved. According to MIC (2011) survey in the 2010 "The users use the most frequently way in online shopping" is group buying. And the utilization rate reached 27 %, compared with 2009 growth of nearly 16 %. The Taiwan group buying market output value reached NT$7.16 billion in 2010, and reached NT$8.95 billion in 2011 [1]. In addition, the foreign group buying market growth also considerable, according to EMarketer [2] reported that the United States group buying forecast revenue to 2015 will be $3.93 billion. It can be seen, whether foreign or domestic, group buying market became more important in the online shopping. Group buying is defined as consumer through open sharing platform on the website, gathered other consumer who have the same demand for products or services and through the way of bulk purchase to enhance their bargaining power, thereby obtaining a product discounts and deals [3, 4].

Group buying recently has been recognized as a kind of social commerce integration, which promotes customers spending together [5]. Sterling et al. [6] point out that local social commerce bring the explosion of group buying, which reflects dramatic growth of group commerce. While group buying can be said is a group commerce, a future online commerce trend can be predicted to group buying and social commerce [7, 8]. Since social commerce is a new online paradigm in the social networking epoch [9], there is an imperative for understanding of the key variables that influence online group buyers' decision-making in social commerce.

With the popularity of social media, such as: Facebook and Twitter, it create a new information delivery platform in e-commerce market, it is called social commerce and its main function is to make use of community strength to conduct a variety of commerce activities on the social networking sites [10]. According to Taiwan's Lotte [11] announced global consumer on online shopping survey found that more than 40 % of Taiwan's consumers prefer to recommended product to friends on social network site, it is much higher than the United Kingdom, Canada, the United States and Japan 20 %. This evident somehow show that the Taiwanese buyers like to share, this also foster social commerce flourishes. In addition, according to the Central News Agency [12] reported that Facebook has become part of many people's lives; the shopping community on the Facebook in the past year have sprung up everywhere. The main reason is that many community fans they are also friends, through the word of mouth and cultivate a loyal customer, they will "one pull one".

14.2 Guanxi in Group Buying

Group buying is the way of group purchase and interpersonal relationships become extremely important. Especially, under the Chinese culture is more emphasis people interaction Guanxi. People tend to through the ren-quin (favors) to express their social friendship and group interpersonal Guanxi. As a result, when your

friends ask you to join the group buying together and the people who affected by the Chinese culture more prone to embarrassed to refuse the mentality. The main reason is to give friends face or ren-quin that affects people's purchasing decisions. For example, people often purchase insurance in order to give the face for others or ren-quin factor. In addition, when friends join and discussion group buying more, people in order to integrate into the circle of friends and more easily affected by friends, thus have conformity behavior.

In the unique Chinese culture environment, people more easily affected by social and arising from conformity behavior or because of ren-quin and face must join group buying. That is, price discounts and free deals are already not the main reason for join group buying, the real impact of consumers to join group buying come from Guanxi between people and involved ren-quin and face. This phenomenon more generally occurs in the close links of social community. Of course, we can know that through social community can improve group buying and driven social commerce to flourish. However, based on the perspective of consumers, whether consumers join group buying it is because that they want to buy the goods or not. Maybe it is possible conformity or in order to show good image in front of public and difficult to refuse ren-quin factor. These consumers participate in group buying influenced by social community. In the past research investigate that the factors affecting group buying tend to focus on the limited time deals or price discounts but few studies focused on group buying behavior under Chinese culture that affected by social impact, face and ren-quin factors. Therefore, this study suggests that the face, ren-quin factors under Chinese cultural and the conformity of compulsive buying behavior research issues is worthy of researchers to in-depth explore in the further. The conceptual framework of this study is shown in Fig. 14.1.

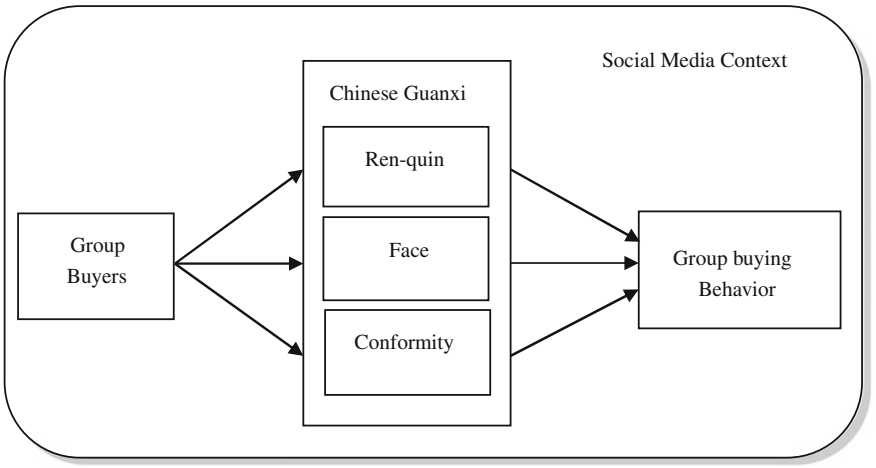


Fig. 14.1 Conceptual Framework of this study

14.3 Pre-Conclusion

14.3.1 Academic Implications

Previous studies of the motivation to participant in group buying have focused on product quality or price and trust for business. But this study investigating the group buying from the perspective of Chinese culture and finding consumers always participant group buying affected by ren-quin and face. Therefore, this study will help researchers to understand that there are different factors to participant group buying in different culture. The Chinese culture interpersonal relationship "Guanxi" is somehow different from the Western factors of trust and commitment, therefore it is worthy of investigate and understanding.

14.3.2 Practical Implications

The Chinese culture indivisibility affects consumers to participant in online group buying. The business should understand this Chinese interpersonal relationships "Guanxi" to accumulate more consumers and to advance group buying market. Therefore, businesses can plan activities and build community to create group cohesion so as to further strengthen consumer with each other. Consumers participant group buying affected by Guanxi and shares the reciprocal with each other. Businesses and customers can get their desired benefit through the mechanism of group buying to creating a win–win situation for both parties.

References

1. Market Intelligence Center (2012) Output value of taiwan e-commerce. http://mic.iii.org.tw/index.asp
2. EMarketer (2011) A bright future for daily deal sites. http://www.emarketer.com/(S(klwlrh45gcyhebv15bgrl255))/Article/Bright-Future-Daily-Deal-Sites/1008283
3. Kauffman RJ, Wang B (2001) 'New buyers' arrival under dynamic pricing market microstructure: the case of group-buying discounts on the Internet. In: 34th Hawaii conference on systems science, Maui, Hawaii
4. Anand KS, Aron R (2003) Group buying on the web: a comparison of price-discovery mechanisms. Manag Sci 49:1546–1562
5. StoreFrontSocial (2012) Group buying—a perfect example of social commerce integration. storefrontsocial.com
6. Sterling G, Moran J, Evans P (2010) Local social commerce: the explosion of group buying, vol 2013. BrightTalk.com
7. Kelt J (2011) The future of group buying and social commerce. CMO Exclusives
8. Huang Z, Benyoucef M (2012) From e-commerce to social commerce: a close look at design features. Electron Commer Res Appl, in press

9. Kim S, Park H (2013) Effects of various characteristics of social commerce (s-commerce) on consumers' trust and trust performance. Int J Inf Manage 33:318–332
10. Liang TP, Turban E (2011) Introduction to the special issue social commerce: a research framework for social commerce. Int J Electro Commer 16:5–14
11. Taiwan's Lotte (2013) Lotte and Facebook together to start the business community. http://www.rakuten.com.tw/info/release/2013/0320.html
12. Central News Agency. (2013). *Shopping on the Facebook has sprung up*. Available: http://tw.news.yahoo.com/如 臨 店 購 貌-臉 書 購 物 異 軍 突 起-013428839.html

Chapter 15
Introspection of Unauthorized Sharing on Social Networking Sites

Cathy S. Lin and Ting-Yi Lin

Abstract The digital content industry is a remarkable market in the information era. Yet the various types of piracy such as software, music, e-book, comic, animation, and movie piracy always threaten the digital content industries and cause enormous losses in revenue. Especially the emerging social networks play a dramatic role in disseminating much unauthorized sharing on the net. Previous studies and practicians have paying close attention in how to diminish piracy. Recently, the experiences from Swiss and Dutch governments have told that there is another way to view piracy. While the sharing behaviours is quite common on social networking context, this study is an introspection of digital content industries and tries to propose the positive viewpoint that drive users eventually spend the money on those digital content products.

Keywords Digital content industry · Piracy · Sharing · Social networking

15.1 Introduction

As technology developed, the digital content industry have begun to develop, such as music, animation, etc., and even the traditional content industry have also dropped into the digital content market. Music is already a kind of well-known digital content product. According to the data of music industry released by International Federation of Phonographic Industry [1], in 2010, the output value in

C. S. Lin (✉) · T.-Y. Lin
National University of Kaohsiung, 700, Kaohsiung University Rd, Nanzih District, Kaohsiung 811, Taiwan, Republic of China
e-mail: cathy@nuk.edu.tw

T.-Y. Lin
e-mail: sunyax711@gmail.com

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_15,

United States is 108.571 billion NT dollars. Output value of music is 22.36 billion NT dollars in Japan, 16.8 billion NT dollars in Taiwan, 3.491 billion NT dollars in South Korea, and 6.93 billion NT dollars in China.

In addition to the music industry, which is worth mentioning is animation industry in Japan. In 2010, the out value of Japanese animation was 407.3 billion NT dollars [2]. The out value of South Korea and Taiwan were 13.1 billion NT dollars and 43 billion NT dollars [3]. In Chinese, the out value is calculated by animation and comics, and it is 224.4 billion NT dollars [4]. Although there lacks a separate calculation data in China, the growing of Chinese animation can be found by comparing the amount of animation of China and Japan. In Japan, the animation industry has developed for several years and its growth has begun to decline, but China's animation industry is gradually growing. From 2007 to 2010, the number of Japanese animation was reduced from 195 to 250, but the Chinese animation grew from 385 to 194 [3].

15.2 Threaten of Piracy

Since there seemed to be a big market of the digital content industry, the network also brought negative effects. Many companies received the impact of piracy, because people could take advantage of convenient way to replicate and spread the digital files in the internet. To deal with this problem, International Federation of Phonographic Industry is committed to the anti-piracy advocacy and piracy prevention in music industry and Business Software Alliance (BSA) also do for the same effort in the software industry. In addition to the product which is the digital content itself, like music or software, the publishing industry had also been a considerable impact. Novels, comic books and other publications which can be scanned into electronic files easily are spreading on the Internet now. And it is a big shock for Japan, which is famous by comics and animations. The comic industry which had a steady growth from 1999 turned to decline after 2005 [5], and also the animation industry [2].

However, faced with the threat of piracy, the content industry in recent years have taken a variety of methods to try to restore such lose. Therefore, the situation of related industries in recent years is not entirely disappointing in fact.

15.3 Coping with a Piracy Environment

15.3.1 Marketing Strategy of Music Industry

In addition to the high degree of copyright protection mechanisms of South Korea, the phenomenon that Korea's and Japan's idol groups have been attracting people in the world is a possible factor to keep the music industry in good growth. They pay

more attention to the relationships with consumers, in addition to just developing the music. By using of social networks, handshake meeting, etc., they can narrow the distance between consumers and enhance consumers' loyalty. For example, Japanese idol group "AKB48", engaging a image which like "the girl who lives next the door", change the high price threshold, switching to low-cost strategy, and let the public have a feeling that they can strive together with the idols. And, to narrow the distance with consumers and artists, the idols of AKB48 are almost picked from the local underground idols which already had some fans of loyalty [6]. These cases illustrate that, in the music industry, in addition to the prevention of piracy, the industry began trying to be closer with their consumers and win the consumers' heart.

15.3.2 Marketing Strategy of Animation Industry

Although the number of Japanese animation declined in these years, the value was still high and grew to 410.9 billion NT dollars in 2011 [2]. It also grew to 45 billion NT dollars in Taiwan [3]. In China, the output value is calculated by animations and comics, and it grew to 297 billion NT dollars [4]. For the animation industry, the real effect to income is not simply affected by production volume. In addition to program sales, the surrounding commercial and other revenues are included. In the book "Japan's Creative, Moe Economy," the author mentioned that Japanese products had tends to be designed to enhance consumers' favour by making people think that the product is close to themselves [7]. This kind of concept means to attract consumers by using the unique of product's cutely, attractively or so on, and this concept can be often seen in the surrounding commercial in animation and comic industry, such as action figure, drop ornament and so on.

15.4 Summary

Recently the government of Swiss and Dutch have conducted studies to explore the impact of piracy download has on society; the report shows that "even in the current situation where piracy is rampant, the entertainment industries are not necessarily losing money". Therefore, we can found that, in the digital content industry, despite the prosecution of piracy, the foreign government has been gradually thinking about how to use the product characteristics and marketing tools to make closer to their consumers. So, this study proposes a conceptual framework as shown in Fig. 15.1 to demonstrate the two kinds of thinking, one is the positive viewpoint on product itself to attract more customers intent to buy it, and the other is the classical viewpoint of ban on piracy that usually lead users remain inclined not to pay for the products. This is a preliminary work that calls for more studies to have a positive way of thinking in piracy and sharing, especially in the social networking context.

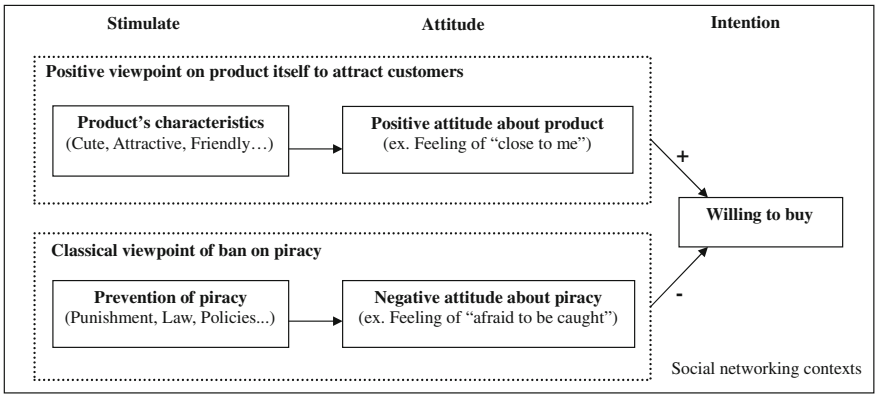


Fig. 15.1 Conceptual framework of this study

References

1. IFPI (2012) Digital music report 2012. http://www.ifpi.org/content/section_resources/dmr2012.html
2. Japan Animation Association (2012) Animation industry report 2012. http://www.aja.gr.jp
3. Taiwan Ministry of Economic Affairs (2011) 2011 yearbook of Taiwan's digital content industry. http://www.dcipo.org.tw/
4. Chinese Culture Media Network (2012) Domestic animation: how to balance yield and output value. http://comic.people.com.cn/BIG5/122400/130243/17140661.html
5. National Institute of Publishing Science Publishing Society (2011) 2011 edition published indicators report. The National Institute of Publishing Science Publishing Society, Tokyo
6. Robinson T (2013) AKB48's plaid skirt economics: amateur idol creative marketing effect. Yuan-Liou, Taipei
7. Sheng K (2009) Japanese's creativity. And MOE economy. Taipei, Business Weekly