Data Mining for Service

Katsutoshi Yada

1 Background of Service Science

In recent years, amidst advancing globalization of the global economy, in both developed and developing countries, the services sector is becoming increasingly important in various fields [2, 9]. In developed countries, service industries comprise a very high percentage of GDP, and even in manufacturing, in order to gain a competitive advantage, there is a focus on services which create added value. Even in developing countries, global companies from developed countries are entering their markets, and service industries are expanding rapidly. With this growing global importance of the services sector, many companies are facing problems in boosting productivity of the services sector.

Service science is one of the most useful approaches for boosting productivity of the services sector [1, 8], which is attracting great attention of people in both business and research. By organically combining tangible and intangible resources in and out of the company, one can define processes which create new value in company activities. Processes which combine resources in this way to create new value are services, and service science scientifically analyzes the dynamics of these processes, and creates new knowledge.

Differing from traditional services research, service science focuses on scientific analysis of processes which create and provide intangible value in services. That is, service science is the cycle of steps described as follows (Fig. 1). By accumulating data which electronically records the status and changes of various subjects and phenomena, and by analyzing these, one can understand phenomena and build models. Then, based on knowledge obtained from these models, one repeatedly improves operations, creates new services, records those results as data, and verifies them. Data mining research has mainly dealt with generating and accumulating data from

K. Yada (✉)
Data Mining Laboratory, Faculty of Commerce, Kansai University, 3-3-35 Yamate-cho,
Suita-shi, Osaka, Japan
e-mail: yada@kansai-u.ac.jp

K. Yada (ed.), *Data Mining for Service*, Studies in Big Data 3,
DOI: 10.1007/978-3-642-45252-9_1, © Springer-Verlag Berlin Heidelberg 2014

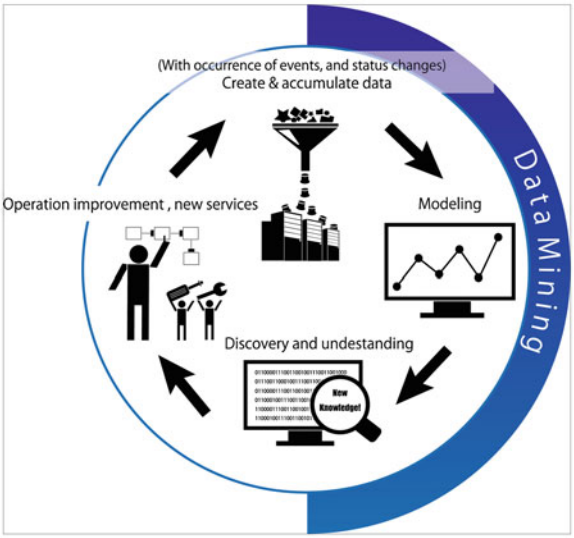
(With occurrence of events, and status changes)
Create & accumulate data
Modeling
Data Mining
Discovery and undestanding
New Knowledge!
Operation improvement , new services

Fig. 1 Cycle of service science

these cycles, then modeling, and finally discovering and understanding new knowledge. The services sector traditionally tended to rely on intuition and experience, but service science introduced data mining technology in research, and is expected to generate great added value.

2 Data Mining for Services

Services have important characteristics which differ from general products [2, 7]. Services do not have a physical form, so one cannot make a large volume of services and store them as inventory. A producer of services generally provides some function directly to the party receiving the service; that is, production and consumption of services occur at the same time. For this process of production and consumption of services, it is difficult to apply a scientific approach which collects data, clarifies causal relationships, does modeling, and verifies improvement proposals. This is because it was traditionally very difficult to record as data the events in which production and consumption of services occur.

In modern business, an important quality of services is the creation of value by complex processes which combine diverse tangible and intangible assets. Many interested parties bring various functions together to produce one service, so when creating services, data is independently generated in diverse situations. Also, those processes which provide services are complex and occur diachronically, so these changes over time have a large impact on the quality of service. Such qualitative

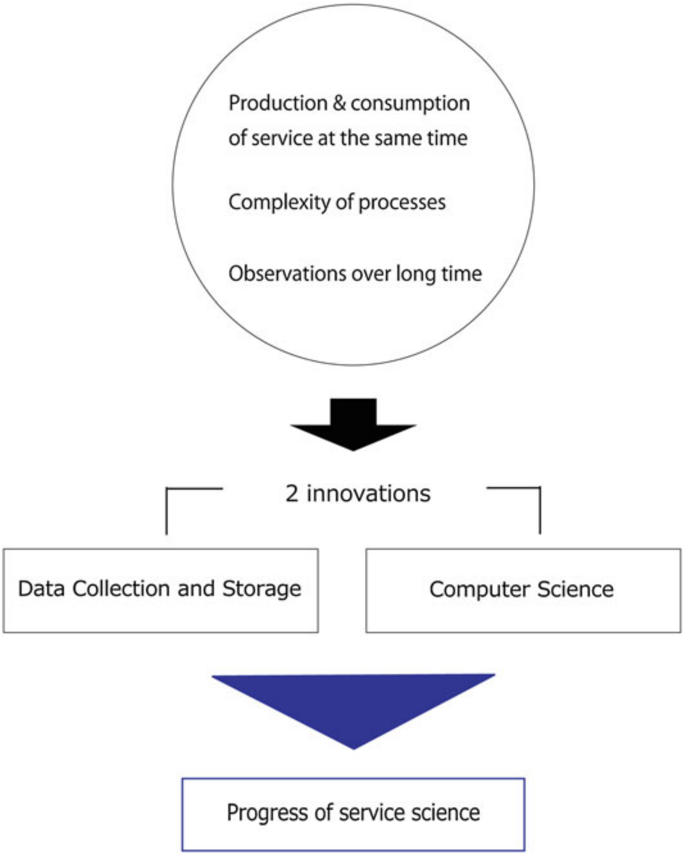
Production & consumption of service at the same time
Complexity of processes
Observations over long time
2 innovations
Data Collection and Storage
Computer Science
Progress of service science

Fig. 2 Two innovations and service science

changes in services create two phenomena which make scientific analysis of services even more difficult. First, one must accumulate data on complex processes [6]; second, changes are recorded diachronically, and large volumes of data must be handled [4].

Figure 2 shows how in recent years, two innovations which arose under these conditions are working to change services industries [10, 11]. First are the dramatic developments in information devices which record data: internet click streams, video monitoring, sensor technologies, etc. This makes it possible to electronically record various service process events which were traditionally difficult to record. Such data contains rich information on the results of events, and also on the processes which create these results. Cutting edge information devices enable automatic accumulation of large volumes of such information.

Second, in the field of computer science, various fundamental technologies for analyzing large-scale diverse data have been developed. It has become possible to

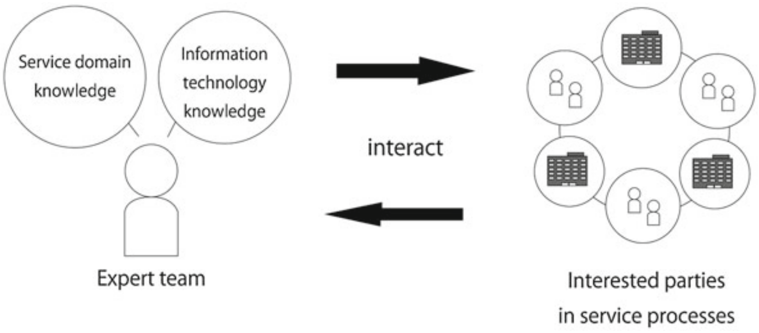
Service domain knowledge
Information technology knowledge
interact
Expert team
Interested parties in service processes

Fig. 3 Mechanism for success in data mining

apply various types of research in computer science fields: distributed processing technologies such as cloud computing that support the accumulation and management of large-scale data, and analysis technologies such as machine learning and advanced statistics to quickly extract useful patterns from large-scale and time-series data. This provides the technical environment to enable relatively cheap and easy analysis of large-scale, complex, time-series data, and conditions to scientifically analyze services.

The two innovations described above have brought large contributions to the field of data mining in the cycle of service science [6], that is, in the generation and accumulation of data, modeling, discovery of new knowledge, and in a deep understanding of phenomena. Along with the phrase “big data”, recent years have seen service science introduce and then greatly develop data mining. However, in this cycle, there are still barriers between finding new knowledge from data, and creating new value and services for customers. Knowing something does not always lead to using that knowledge to create a new value or service. In order to create a new value or service, one must repeat the above cycle at the actual service site, and continually create incremental change. However, continually creating such new changes is very difficult work. This is because to repeat this cycle and continually create changes, one must have domain (specialized) knowledge in that applied field, and also expertise in information technology; such human resource skills are extremely rare.

As far as we know, successful examples of data mining research and practice in the services field share a common pattern (see Fig. 3): Unceasing changes are created when an individual or specialist team has domain knowledge in the services field to which this data mining is applied, is also well versed in information technology, and they repeatedly interact with interested parties involved in the service processes. As described above, production and consumption of services occur at the same time, so people who have domain knowledge and are aware of the unique problems in that field must improve existing algorithms and develop new technologies which suit these characteristics. Therefore, to skillfully utilize data mining in the services field, one must understand both the unique characteristics of the application field, and the applied technologies. However, in traditional service science and data mining research, there is no research which clarifies the differences in each field's unique domain knowledge and characteristics as well as in the applied algorithms.

This book covers the fundamental technology for data mining as well as cases of application in diverse service fields, and thereby clarifies the characteristics of data mining applications in services fields. In order to understand data mining for services, this book focuses on two points. First, by covering cases of applications in diverse service fields, we study various aspects of data mining applications for services. For the data mining described above to be introduced in services, one must develop unique algorithms and techniques based on domain knowledge of those services. Therefore, to introduce data mining and establish efficient services, one must understand the characteristics of the services and the suitability of applied algorithms and techniques. This book describes cases of applications in diverse services fields such as a service for searching research papers, and use in marketing of social networking services; it thereby provides a deep understanding of data mining applications in services.

Second, this book focuses on diverse data generated and accumulated in various processes of services, and their combination [3, 5, 12]. New information devices such as sensors and video devices enable data to be accumulated on events which occur in various situations of services, and that data comes in various types, structures, and sizes. Combining these and extracting patterns and knowledge unique to a service is difficult. This book covers data ranging from structured transactions data to unstructured conversation data in social networking services, and clarifies how diverse data is combined and how data mining contributes to these services.

3 The Organization of this Book

The organization of this book is as follows. This book is comprised of four parts, each containing multiple chapters. Part I, "Fundamental Technologies Supporting Service Science", describes research on the latest algorithms and techniques which support data mining in the services sector. Diverse services require various algorithms and techniques which suit each service, and sometimes new combinations of these must be created. This part describes various fundamental technologies used in services, such as clustering, feature selection, and dimensionality reduction.

For example, selecting important features from a huge volume of features is one important task in data mining. Chapter, "Feature Selection Over Distributed Data Streams" discusses mainly feature selection techniques, especially focusing on techniques to select features from data groups accumulated in distributed systems. Chapter, "Learning Hidden Markov Models Using Probabilistic Matrix Factorization" discusses the Hidden Markov Model (HMM), which estimates path parameters from time series observation data. HMM uses an EM algorithm to estimate parameters, but calculations take a long time. This chapter describes a technique which uses efficient probabilistic matrix factorization (PMF). Chapter, "Dimensionality Reduction for Information Retrieval Using Vector Replacement of Rare Terms" discusses dimensionality reduction, which is a core technology in text mining. When trying to extract useful characteristics from a huge volume of text data, dimensionality reduction, which narrows characteristics down to important key words, is one of the most important technologies. This chapter describes a dimensionality reduction technique which uses rare terms and co-occurring terms. Chapter, "Panel Data Analysis Via Variable Selection and Subject Clustering" studies clustering techniques for panel data used in marketing and economics. This chapter proposes a technique which does variable selection and clustering using BIC at the same time, and describes results when this is applied to U.S. company survey data.

Part II, "Knowledge Discovery from Text", describes research on knowledge discovery from huge text data. One can create various new services from huge amounts of text information, such as Internet comments, call center conversation recordings, or an academic database. This part describes cases in which text mining was applied to services, and explains the core technologies used.

Chapters, "A Weighted Density-Based Approach for Identifying Standardized Items that are Significantly Related to the Biological Literature", "Nonnegative Tensor Factorization of Biomedical Literature for Analysis of Genomic Data" and "Text Mining of Business-Oriented Conversations at a Call Center", also in Part II, discuss research on knowledge discovery from biological literature. Chapter, "A Weighted Density-Based Approach for Identifying Standardized Items that are Significantly Related to the Biological Literature" studies techniques using abstracts of published biological literature to classify their content. The authors propose a weighted density-based approach, and found that this can be used to classify papers more accurately than previous techniques. Chapter, "Nonnegative Tensor Factorization of Biomedical Literature for Analysis of Genomic Data" discusses a more specific issue, by describing a technique using Non-negative Tensor Factorization (NTF) to extract relationships between genes and transcription factors. Chapter, "Text Mining of Business-Oriented Conversations at a Call Center" describes a case in which text mining was applied to data on conversations between operators and customers in a call center. The authors present a framework using a Support Vector Machine (SVM), and provide a technique which extracts conversational expressions and words which affect the customer's situation.

Part III, "Approach for New Services in Social Media", contains research that apply data mining to data accumulated in Internet social networking services which have grown rapidly in recent years, to create new knowledge and services. Chapter, "Scam Detection in Twitter" discusses scam detection from Twitter. This chapter proposes a scam detection technique using suffix trees which combine Twitter's unique data content, where one tweet is a short sentence using a special symbol such as "#". Chapter, "A Matrix Factorization Framework for Jointly Analyzing Multiple Nonnegative Data Sources" describes a search method using combined data from multiple social media, applies this to data from YouTube and BlogSpot, and shows its usefulness. Social networking services providing various services exist, but it is also possible to combine these to create other new services. Chapter, "Recommendation Systems for Web 2.0 Marketing" proposes a framework which combines multiple data sources, and applies them to a recommendation system. It is difficult for the information needed for collaborative filtering to be gathered only in a company's own Internet shop search data. This chapter proposes a framework which combines product search information and social media information.

Part IV, "Data Mining Spreading into Various Service Fields", describes leading research fields in which data mining is applied to diverse data obtained from new information devices, with various services being created. Chapter, "Handling Imbalanced and Overlapping Classes in Smart Environments Prompting Dataset" describes a case where data mining is applied to smart home environments based on home electronic products connected to networks, a case which has spread rapidly in recent years. Data created in actual events often contains imbalanced data, such as mistaken use of home electronics. This chapter proposes ClusBUS, a clustering technique to handle the overlapping class problem created by imbalanced data. Chapter, "Change Detection from Heterogeneous Data Sources" studies a technique of change detection, focused on sensor data which has been attracting the most attention as big data in recent years. The authors describe the singular spectrum transformation technique for change-point detection, which handles dynamic characteristics in actual heterogeneous sensor data, and shows its usefulness. Chapter, "Interesting Subset Discovery and Its Application on Service Processes" describes applied cases of data mining using various databases accumulated in companies. It shows specific cases where data mining is applied to databases which mix continuous values and discrete values: employee satisfaction surveys, IT support system failure surveys, data processing performance surveys for data centers, etc. Chapter, "Text Document Cluster Analysis Through Visualization of 3D Projections" concludes with problems of visualization, which most affect the performance of data mining in business applications. This chapter shows a framework and system to visualize the clustering process and results in text mining, so non-technical users can also understand the phenomena.

References

1. Chesbrough, H., Spohrer, J.: A research manifesto for service science. Commun. ACM **49**(7), 35–40 (2006)
2. Fisk, R.P., Grove, S.J., John, J.: Interactive Services Marketing, 3rd edn. Houghton Mifflin Company, Boston (2007)
3. Hamuro, Y., Katoh, N., Yada, K.: MUSASHI: flexible and efficient data preprocessing tool for KDD based on XML. In: DCAP2002 Workshop Held in Conjunction with ICDM2002, pp. 38–49 (2002)
4. Hui, S.K., Fader, P.S., Bradlow, E.T.: Path data in marketing: an integrative framework and prospectus for model building. Mark. Sci. **28**(2), 320–335 (2009)
5. Ichikawa, K., Yada, K., Washio, T.: Development of data mining platform MUSASHI towards service computing. In: 2010 IEEE International Conference on Granular Computing (GrC), pp. 235–240 (2010)
6. Ohsawa, Y., Yada, K. (eds.): Data Mining for Design and Marketing. CRC Press, USA (2009)
7. Sasser, W.E., Olsen, R.P., Wyckoff, D.D.: Management of Service Operations, Allyn and Bacon, Boston (1978)
8. Spohrer, J., Maglio, P.P., Bailey, J., Gruhl, D.: Steps toward a science of service systems. IEEE Comput. **40**(1), 71–77 (2007)

9. Tukker, A., Tischner, U.: Product-services as a research field: past, present and future. Reflections from a decade of research. J Cleaner Reprod. **14**(17), 1552–1556 (2006)
10. Yada, K., Washio, T., Koga, H.: A framework for shopping path research. In: Proceedings of SIAM International Workshop on Data Mining for Marketing, pp. 31–36 (2011)
11. Yada, K.: String analysis technique for shopping path in a supermarket. J. Intell. Inf. Syst. **36**(3), 385–402 (2011)
12. Yada, K., Hamuro, Y., Katoh, N., Washio, T., Fusamoto, I., Fujishima, D., Ikeda, T.: Data mining oriented CRM systems based on MUSASHI: C-MUSASHI. In: Tsumoto S. et al. (eds.) Active Mining, LNAI 3430, pp. 152–173 (2005)