The Electronic Library

Exploring topics related to data mining on Wikipedia

Yanyan Wang, Jin Zhang,

Exploring topics related to data mining on Wikipedia

Yanyan Wang and Jin Zhang
School of Information Studies, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Received 22 September 2016
Revised 16 March 2017
Accepted 9 April 2017

Abstract

Purpose – Data mining has been a popular research area in the past decades. Many researchers study data-mining theories, methods, applications and trends; however, there are very few studies on data-mining-related topics in social media. This paper aims to explore the topics related to data mining based on the data collected from Wikipedia.

Design/methodology/approach – In total, 402 data-mining-related articles were obtained from Wikipedia. These articles were manually classified into several categories by the coding method. Each category formed an article-term matrix. These matrices were analysed and visualized by the self-organizing map approach. Several clusters were observed in each category. Finally, the topics of these clusters were extracted by content analysis.

Findings – The articles obtained were classified into six categories: applications, foundation and concepts, methodologies, organizations, related fields and topics and technology support. Business, biology and security were the three prominent topics of the applications category. The technologies supporting data mining were software, systems, databases, programming languages and so forth. The general public was more interested in data-mining organizations than the researchers. They also focused on the applications of data mining in business more than in other fields.

Originality/value – This study will help researchers gain insight into the general public's perceptions of data mining and discover the gap between the general public and themselves. It will assist researchers in finding new techniques and methods which will potentially provide them with new data-mining methods and research topics.

Keywords Social media, Data mining, Social Web mining, Theme discovery

Paper type Research paper

Introduction

With the development of internet technologies, information and data are produced, shared, and stored much faster than before. The volume of data grows every day as companies capture large amounts of data about markets, products, customers and suppliers. Individuals also receive large quantities of data from their daily life and the internet. Moreover, the evolution of mobile devices, social media and Web technologies boosts the growth of data and information. It is difficult, however, to deal with huge data sets using traditional data analysis approaches. Because of these circumstances, the concept of data mining was created.

To explore the internal relationship and patterns of data, data mining was proposed in the 1990s. Since then, data mining has been studied and used as a useful research method. As more and more people face the problems of data analysis and management, this concept has been widely accepted and the related techniques and methods have been frequently used by both researchers and general users. Research topics about data mining can be found in a large number of publications. In addition, there are introductions and discussions of data mining on the internet, especially on social media platforms. Different from data mining research studies, the content of data mining on social media platforms has its own features.

The Electronic Library
Vol. 35 No. 4, 2017
pp. 667-688
© Emerald Publishing Limited
0264-0473
DOI 10.1108/EL-09-2016-0188

Since the use of data-mining theories, methods, and technologies continually increases, it is necessary to gain insight into data-mining and related topics. Previous research papers have studied various aspects of data mining, but few have explored the data-mining-related topics based on data collected from social media. Because Wikipedia is the largest online knowledge collaboration, to fill the gap, this study aims to explore the data-mining-related topics on Wikipedia. The self-organizing map (SOM) approach, a machine learning approach, was applied to this data analysis.

Literature review

Data mining

Data mining is a method to reveal previously unknown and reliable insights from large data sets (Elkan, 2001). Because the massive volume of data from different fields keeps growing, useful analysis methods and techniques are urgently needed. Therefore, data mining has become an increasingly important research area (Liao *et al.*, 2012).

With the development of data mining, a variety of methods and techniques from other areas have been introduced to the data-mining area, such as classification, clustering and database technology (Liao *et al.*, 2012). In Han and Kamber's (2006) book, they pointed out the disciplines that most influence and improve the data-mining method. These are statistics, machine learning, database systems, warehousing and information retrieval. Meanwhile, data mining has impacted other research fields, such as chemistry, medicine, business and so forth (Aljumah *et al.*, 2013; Borghini *et al.*, 2010; Zhang *et al.*, 2013).

In addition to prediction, data mining has other functions. Han and Kamber (2006) summarized the different patterns that can be mined: frequent patterns, associations and correlations; classification and regression; clustering analysis; and outlier analysis. Fu (2011) gave a similar opinion on time series data mining, which said that the main tasks of time series data mining are pattern discovery and clustering, classification, rule discovery and summarization. Different data-mining methods and techniques have been proposed and applied to accomplish different tasks. For example, k-means, fuzzy c-means and SOM are frequently used in clustering analysis. Moreover, there are specific methods to mine certain types of data, like the model-based sequence clustering methods for mining temporal data (Law and Kwok, 2000).

Social Web mining

Web mining, as a branch of data mining, is gradually playing increasingly important roles in research. Social Web mining is one of the primary components in studies related to Web mining and social media. Social media is the way people generate, share and communicate information in virtual communities and networks (Ahlqvist *et al.*, 2008). Under the big umbrella of social media, social media sites and applications vary a lot. For instance, Twitter, which is regarded as a microblog, allows users to communicate and create posts of less than 140 characters (Kwak *et al.*, 2010), while Wikipedia provides opportunities for collaborative information and knowledge production (Bruns, 2006). With the development of mobile devices, geo-mapping tools (e.g. Google Maps) and self-tracking applications (e.g. Quantified Self) have been invented.

The methods applied in social Web mining can be classified into two groups: social network analysis methods and sentiment analysis methods. Social network analysis tries to reveal human relationships and connections (Hansen *et al.*, 2010). In recent years, various tools have been invented to analyse and visualize social networks, such as UCINET, Pajek, NetworkX in Python and igraph in R (Borgatti *et al.*, 2002; Kolaczyk and Csárdi, 2014; de Nooy *et al.*, 2011). Sentiment analysis is known as *opinion mining*, which is related to *text mining* (Thelwall *et al.*, 2011). Therefore, methods for text mining are also used in sentiment analysis. Sentiment analysis aims to predict opinion or emotion from content by consistent, repeatable, algorithmic approaches (Hill *et al.*, 2013). A number of tools are available for extracting opinions and emotion, like the SAS Text Miner, Python and R (Miner, 2012; Paltoglou, 2014).

Research on Wikipedia

The concept of *crowdsourcing* was first defined by Howe (2006) as a Web-based business model. Over the past 10 years, many crowdsourcing platforms have emerged, such as Yahoo! Answers, ESP game, Wikipedia, Mechanical Turk-based systems and so forth (Doan *et al.*, 2011). With the creation of these platforms, the definition of crowdsourcing has evolved, and nowadays, crowdsourcing platforms are regarded as online distributed problem-solving and production systems which allow a huge number of humans to contribute (Brabham, 2008; Doan *et al.*, 2011). The users on crowdsourcing platforms have their own characteristics:

- they are both users and producers;
- the number of contributors is flexible;
- the engagement of users is not restricted by location; and
- the collaboration between contributors is informal and loose (Benkler and Nissenbaum, 2006; Bruns, 2007).

As a typical crowdsourcing platform, Wikipedia is a text-based system with relatively low richness of medium compared with content communities such as YouTube (Kaplan and Haenlein, 2010). Wikipedia only allows simple interactions among contributors, which distinguishes it from social network sites, Facebook, for example. Therefore, the relationships among Wikipedia contributors are simpler than the relationships among the users on other social network sites. Another feature of Wikipedia is that, although personal identity is core to some social media platforms (e.g. blogs and Instagram), Wikipedia is not a platform for showing personality or building personal image (Kaplan and Haenlein, 2010; Kietzmann *et al.*, 2011). Instead, it is an online encyclopaedia focusing on specific content domains.

The statistics offered by Wikipedia present that it has more than 900,000 content creators generating nearly five million articles. Although some researchers doubted that Wikipedia is not accurate or reliable, it still plays an important role in knowledge sharing and research (Hu *et al.*, 2007). Researchers have proposed measurements to evaluate the quality of Wikipedia articles and have tried to evaluate the contribution value of Wikipedia (Hu *et al.*, 2007; Zhao *et al.*, 2013).

Some researchers have investigated the factors that influence content creation on Wikipedia. These factors include redundancy, polarity and cultural bias (Callahan and Herring, 2011; Moskaliuk *et al.*, 2012). Studying user behaviours on Wikipedia is another primary research topic. Asadi *et al.* (2013) discovered the motivating and discouraging factors for Wikipedians, such as gaining identity and reputation (motivating factors) and copyright violation (discouraging factors). Two types of editors with different editing patterns were proposed: the mediators and the zealots (Iba *et al.*, 2010). Laniado *et al.* (2012) analysed the emotions expressed by editors in talk pages. They found that women tend to participate in discussions with a more positive tone than men, and that administrators are more positive than non-administrators.

There are hundreds of entries and articles relevant to data mining on Wikipedia. Although research about data mining and social Web mining covers various topics, there are few articles studying the data-mining-related topics and themes from the general public's perspective. To fill the gap, this paper aims to explore the main topics and themes of data mining based on articles collected from Wikipedia.

Research methodology

Statement of primary objectives

The primary objective of this study is to explore data-mining-related topics on Wikipedia. The research problems are:

RQ1. What are the main categories of data-mining-related entries on Wikipedia?

RQ2. What are the main topics of each category?

RQ3. What are the similarities and differences between the general public and researchers in terms of their focuses on data mining?

Article selection and categories

Wikipedia is one of the biggest online information collaborations, which is less expertly compared with some other encyclopaedias, such as Citizendium (Flanagin and Metzger, 2011). Citizendium vets the user-generated content before it is posted, while Wikipedia allows anyone to anonymously create and edit entries without inspection. Another example is Nupedia, the immediate predecessor project of Wikipedia. Only academicians can write articles on this platform, and 12 articles were produced after 18 months (Bruns, 2008). The content production on Wikipedia is much faster than Nupedia, which reveals that the general public's contributions on Wikipedia are huge. In addition, different from the encyclopaedias focusing on specific disciplines (e.g. Encyclopedia Dramatica), Wikipedia is a general reference work that provides basic knowledge of certain concepts. Wikipedia users are the general public instead of experts in specific disciplines. According to Bryant *et al.*'s (2005) findings, users determine the value of Wikipedia entries. Therefore, although both the general public and experts generate content on Wikipedia, the contents on this platform reflect the general public's interests. The data-mining-related articles on Wikipedia also show the general public's focus on data mining.

The data-mining entry on Wikipedia was selected as a seed, which is the starting point for data collection. Therefore, this article was regarded as the first-level article in this study. In the "See also" section of the "Data mining" article, there were several links connecting to related articles. These articles are called second-level articles. In this way, third-level articles were also examined. After investigating all the articles, first- and second-level articles were all associated with data mining, while some third-level articles were irrelevant. Meanwhile, some of the articles were duplicated. Therefore, when collecting data, the duplicated articles and the irrelevant articles on the third level were excluded. Fourth-level articles were also investigated, but most of them were not related to data mining. Thus, this level was excluded in data collection.

Every article on Wikipedia refers to one or several synonymous entries. Each article contains several sections, such as content, main text, reference and so on. Although not all articles consist of the same sections, certain sections turn up in almost every article. They are entry/entries associated to the article, content, main text, see also and reference. In this study, only texts from the content and the main body of an article were collected, while sections like "References", "External links", "Bibliography" and "Further readings" were excluded.

A coding method was used to analyse the articles obtained. This method allows researchers to develop concepts and categories from data (Pandit, 1996). The coders read through the articles obtained to compare for similarities and dissimilarities. Similar articles were grouped together, while those containing more different properties were separated.

Consequently, all the articles were assigned into several categories by two trained coders. The Cohen's kappa inter-coder reliability of the coding results is 0.749, larger than 0.7. According to the criteria proposed by Viera and Garrett (2005), the result of the Cohen's kappa agreement test indicates a substantial agreement between the coders.

Matrices processing

The articles were grouped into several categories so that each category contained a set of related items. The frequencies of all the terms in each article were counted. In this paper, a term was defined as a unique word, and stop words were excluded. In addition, the obtained words were stemmed so that the words with the same root were combined into one term, and their frequencies were added together. Each category forms an article–term matrix where its columns are the terms, and its rows are the articles. All the articles were ranked alphabetically, and numbered from 1 to m. A matrix is presented in equation (1). As this equation displays, the matrix has m rows and n columns. The value of each cell (a_{ij}) in the matrix represents the frequency of term j in article i. The matrices obtained are the SOM input matrices in data analysis.

$$M = \begin{pmatrix} a_{11} & a_{12} & \cdot\cdot & .. & a_{1n} \\ a_{21} & a_{22} & \cdot\cdot & .. & a_{2n} \\ .. & .. & a_{ij} & .. & .. \\ & & .. & & \\ a_{m1} & a_{m2} & .. & .. & a_{mn} \end{pmatrix} \qquad (1)$$

Self-organizing map approach

In this study the matrices obtained are defined in a high-dimensional vector space. In every matrix, the terms describe the characteristics of the corresponding articles. The high-dimensional vector space contains the connections among terms. The SOM approach is capable of transforming high-dimensional data to low-dimensional SOM display, and still keeping the characteristics of articles and the connections among terms. Therefore, this approach is suitable for this study. Meanwhile, SOM is a visualization tool that illustrates the similarities between items in its output. This feature offers researchers more straightforward results compared with other clustering methods.

First introduced by Kohonen (1990), the SOM approach is a popular unsupervised learning approach. It is a widely used neural network method that measures similarities between items of input data so as to form similarity graphs. The SOM approach has been applied in various research fields, such as finance, industry, biology and so on (Bernstein *et al.*, 2013; Deboeck and Kohonen, 2013; Fuertes *et al.*, 2010; Sarlin *et al.*, 2012). In the field of library and information science, this approach is usually applied to document cluster analysis and to information-retrieval algorithm, and to explore and extract information from documents (Lin *et al.*, 1991; Suchanek *et al.*, 2009; Zhang, 2007). An *et al.* (2011) used the SOM approach to cluster journals in library and information science in terms of their indexed subjects and obtained 19 clusters for the 60 investigated journals. The SOM approach plays an important role in these research studies, which shows its usefulness and effectiveness in text data clustering.

The whole procedure of SOM is a recursive regression process (Kohonen *et al.*, 2000). A SOM contains many nodes, which are also known as neurons. The number of neurons on a SOM is determined by users. The nodes are organized on a low-dimensional grid. Each node stands for an n-dimensional weight vector, and n equals the number of columns (Himberg *et al.*, 2000).

Several toolboxes in MATLAB contain the functions of SOM, such as the neural network toolbox. In this study, the SOM toolbox created by Himberg *et al.* (2000) was used. This toolbox provided functions for data input, normalization and training, as well as SOM creation and visualization. Multiple algorithms and visualization functions have been proposed for SOM. In this case, the batch algorithm and the cell visualization function were applied. They reveal the distribution of the items in each matrix on the SOM display. Items with shorter distances among them are more similar than those with longer distances. Moreover, the similarity between items could also be indicated by the colour of a SOM output. The colour projected to the SOM display background is determined by a unified distance matrix (U-matrix) (Ultsch and Siemon, 1990). Higher values of the U-matrix stand for cluster borders, while lower values represent clusters.

According to the distances between items and the colours in the background, the items of each matrix were clustered. The criteria for clustering the items was:

- the items located in the same SOM node were grouped into one cluster; and
- items located in two or more nodes, but the nodes were adjacent or separated by only one empty node, were grouped into one cluster.

The Organizations category contained much fewer items than the other categories; therefore, only the first criterion was applied to cluster the items in the Organizations category.

After the clusters of each matrix were obtained, the topics of each cluster were extracted from the content of the items in it. Because the items in this study are the Wikipedia entries, the researcher reviewed the full text of these entries to identify their topics. One cluster is allowed to have multiples topics, while different clusters may have the same topics.

Results and discussion

Categories and matrices

Based on the data collection strategy, a total of 402 articles related to data mining on Wikipedia were selected. Among these 402 articles, there were 45 second-level articles and 357 third-level articles. The 402 articles were assigned to six categories by the coding method, and the coding scheme is presented in Table I. According to this coding scheme, 69 articles were assigned to the Applications category, while 77 articles were allocated to the Foundation & Concepts category. The Methodologies category contained 88 articles, the Organizations category included 28 articles and the Related Fields & Topics category had 53 articles. The remaining 87 articles were grouped into the Technology Supporting category. These six categories indicate that Wikipedia editors and users focus on these themes of data mining: applications, related theories and concepts, methodologies, organizations, fields and topics relevant to data mining and technologies that support data mining.

As stated before, an SOM input matrix was constructed for each category. Accordingly, six article–term matrices were created in this study. The size of each matrix is listed in Table II. It is obvious that the Applications matrix (*AM*) is the largest one among all of the six matrices, and it includes 69 articles and 2,046 terms. The Methodologies matrix (*MM*) and Technology Supporting matrix (*TSM*) have similar sizes, and rank second and third, respectively. The Foundation & Concepts matrix (*FCM*) occupies the fourth position and the Related Fields & Topics matrix (*RFTM*) and Organizations matrix (*OM*) are the smallest two matrices among the six matrices. In general, the categories containing more articles have larger matrices.

Six SOM displays were obtained based on the six matrices and the SOM technique. Because the sizes of these matrices varied a lot, it was rational to process different matrices in different map sizes. Furthermore, the size of every SOM display was adjusted repeatedly

Category	Definition	Examples
Applications	Articles about the applications of data mining	Business intelligence, Mass surveillance, Drug discovery
Foundation & Concepts	Articles about the theory, foundation, and concepts of data mining	Analysis, Data fusion, Extract, Transform, Load
Methodologies	Articles about the methods, approaches, and algorithms used in data mining	Cluster analysis, Odds algorithm, Statistics
Organizations	Articles about programs, projects, companies, agencies, societies, communities, and universities related to data mining	Bioinformatics companies, National Security Agency, Geodi
Related Fields & Topics	Articles about related fields and topics, and concepts, applications, and technologies of them	Cognitive science, Privacy law, Text mining
Technology Supporting	Tools, techniques, software, systems, databases, and languages used in data mining	Decision tree, List of emerging technologies, SQL

Table I.
Coding scheme

Matrix	Size
Applications matrix	69 × 2046
Foundation & Concepts matrix	77 × 1512
Methodologies matrix	88 × 1624
Organizations matrix	28 × 1375
Related Fields & Topics matrix	53 × 1396
Technology Supporting matrix	87 × 1673

Table II.
Size of each matrix

so as to achieve the most interpretable results. Consequently, the sizes of the SOM display maps obtained from the six matrices are different from each other. Figures 1 to 6 illustrate the SOM displays of the six categories. The coloured bars on the right side of the figures represent different values of the U-matrix. Lower value means higher similarity. In the maps, every number stands for an article, and every circle represents a cluster. Grey circles stand for large clusters, and purple circles represent small clusters. The articles and their serial numbers are listed in Tables III to VIII, and the articles in the same pair of square brackets form small clusters.

Applications matrix

The AM includes the term frequencies of the articles in the Applications category. Its SOM display is demonstrated in Figure 1. According to the clustering criteria, the 69 articles in the category are grouped into three large clusters and a group named "Other". The "Other" group contains some isolated items and small clusters that contain less than four items. In Figure 1, there are three clusters: C1, C2 and C3. C1 and C2 are both located in the blue area. It means these two clusters are quite similar. The articles of each cluster are listed in Table III.

After investigating the articles in C1, it can be seen that C1 is related to applications in business intelligence (e.g. business analytics) and biology (e.g. computational genomics). C2 includes articles relevant to applications in chemistry (e.g. cheminformatics), agriculture (e.g. data mining in agriculture) and business management, especially sales and customer

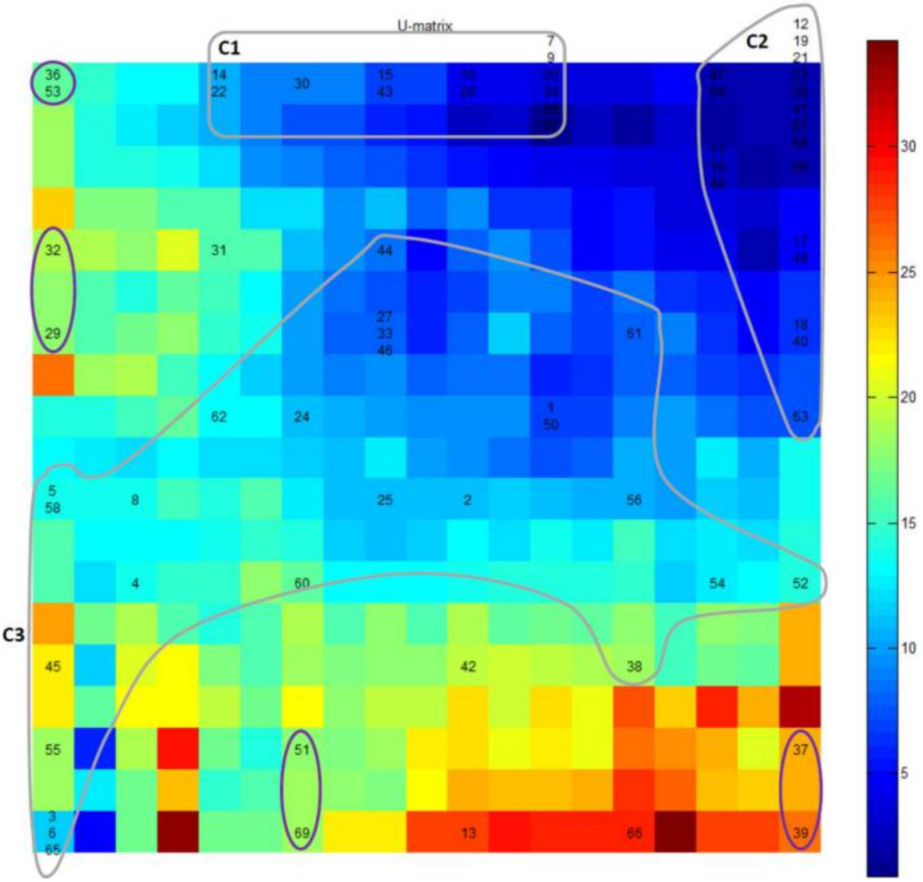

Figure 1.
The Applications matrix display

management (e.g. customer analytics and sales intelligence). These two clusters reflect that the applications of data mining cover multiple areas and disciplines. Most of the articles of the third cluster (C3) are associated with analytics (e.g. learning analytics) and artificial intelligence (e.g. neural networks). There are some unique applications in the three clusters which are the applications in meteorology, onomastics, education, news analytics and so on. One of the small clusters in the "Other" group is associated with the applications in surveillance (e.g. mass surveillance and mass surveillance in the USA). Another small cluster is related to knowledge discovery and management (e.g. Expert system and Knowledge discovery).

The results demonstrate that data mining is applied in a variety of areas and disciplines. Business (14 entries), biology (9 entries) and security (7 entries) are three prominent application areas of data mining in terms of the investigated entries. For instance, according to the Wikipedia article, bioinformatics is an interdisciplinary field which applies data-mining methods in biology. IBM developed the criminal reduction utilizing statistical history system to predict the places of future crimes. On Wikipedia, the applications in business are discussed more than the applications in any other areas. It implies that the general public is more interested in business than in other fields.

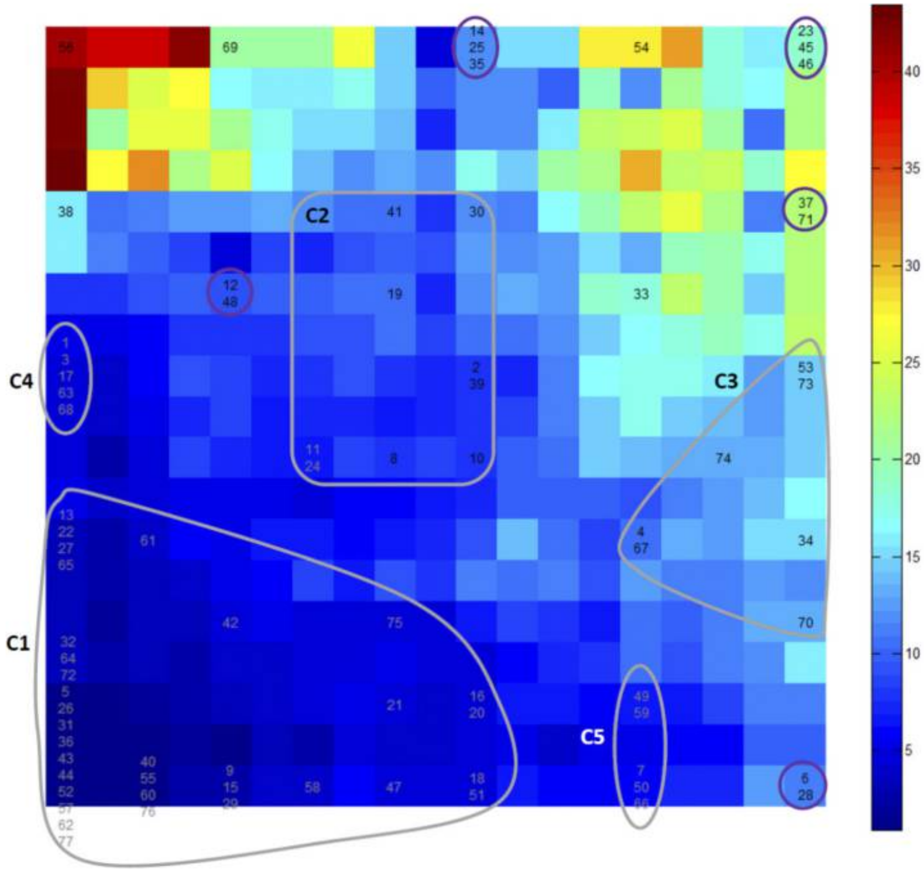

Figure 2.
The Foundation & Concepts matrix display

Foundation & Concepts matrix

The *FCM* contains the term frequencies of the articles about theories and concepts of data mining. Its SOM display is illustrated in Figure 2. There are five large clusters in this figure and all of them located in the blue area. Among them, C1, C2, C4 and C5 are located in the relatively darker area so that they are more similar than the other clusters. Table IV contains the articles in each cluster.

In C1, C2 and C4, there are some theories (e.g. AI effect) and basic concepts (e.g. data visualization, profiling and computational science) about data mining. C3 and C5 include concepts relevant to data-mining methods, such as fuzzy logic, schema matching and pattern recognition. Similar to the "Other" group of *AM*, the "Other" group of this category also includes several small clusters. In this group, some articles are about identity and privacy (e.g. privacy and digital identity), which means that data mining also connects to information security and social security.

Methodologies matrix

The Methodologies matrix (MM) contains the term frequencies of the articles about methodologies utilized in data mining. Figure 3 illustrates the SOM display of the *MM*. There

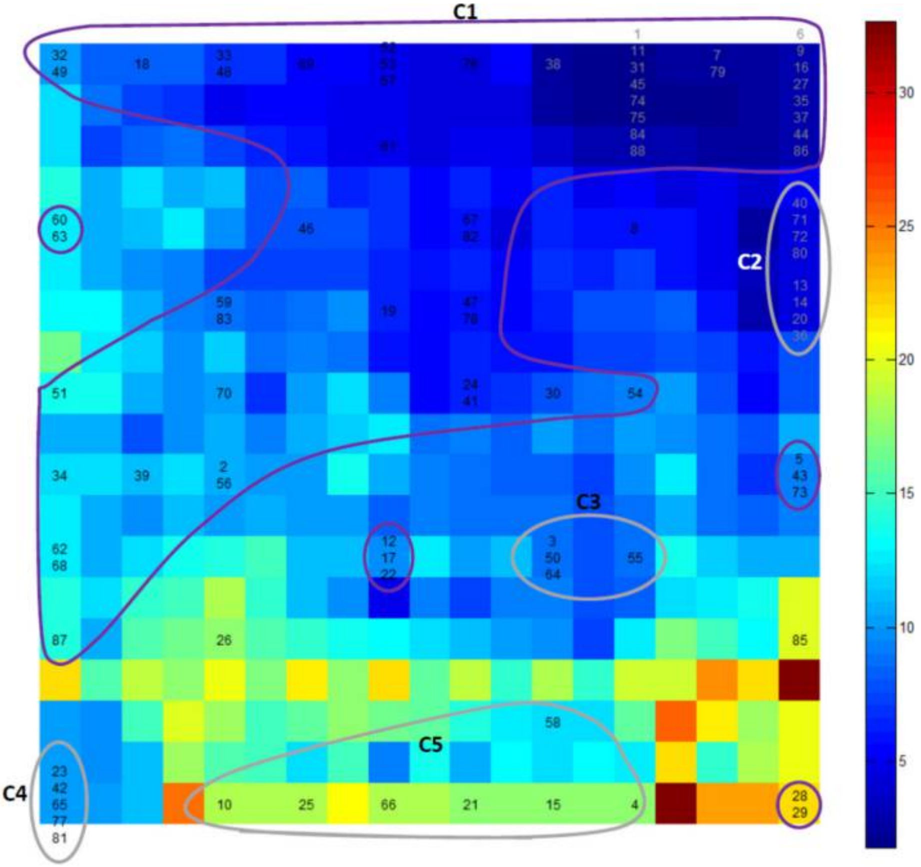

Figure 3.
The Methodologies matrix display

are five clusters in Figure 3, and the remaining items are clustered in the "Other" group as with the previous matrices. Regarding the clusters obtained, all of them are located in blue or green areas which are close to each other. It means these clusters are similar. Table V displays all the items in each cluster.

Investigation of all the clusters and the "Other" group shows that mathematical methods and statistical methods occur in every group. In C3 and C4, all the items are related to mathematics or statistics. Furthermore, in the other three clusters and in the "Other" group, a majority of the items are associated with mathematics or statistics (e.g. Anscombe's quartet, higher-order singular value decomposition, segmented regression and cluster analysis). This indicates that a large part of methods used in data mining stem from mathematics and statistics. Apart from these methods, most of the remaining items have relations to artificial intelligence and neural network (e.g. Kernel machines and instance-based learning). The remaining items in this category are relevant to other fields, such as biology. BioCreative and Universal Darwinism are theories rooted in the biological field.

Looking into the largest cluster in the map, some particular items have a relation to basic theories, such as network theory, morphological analysis and test methods. These methods are popular in many areas, such as psychology, biology and building science (Brum *et al.*,

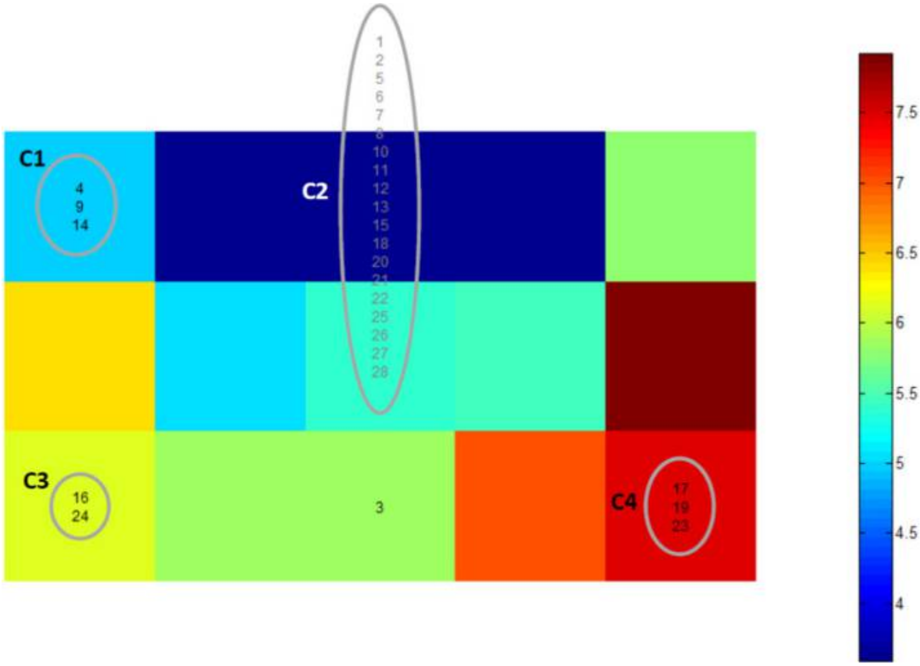

Figure 4. The Organizations matrix display

2013; Johansson *et al.*, 2014; Westaby *et al.*, 2014). Another special item, JXTA, is a peer-to-peer protocol, which is usually adopted in computer engineering (Spaho *et al.*, 2013). Recently, to better present the data-mining results, visualization methods and tools have been widely applied in data mining. In this cluster, multidimensional scaling and parallel coordinates are two items associated with visualization.

Regarding the "Other" group, there is a small cluster in this group which includes two items: gene expression programming and genetic algorithm. These two items both stem from genetics. This again shows that the biological field generates some data-mining methods. Gene expression programming was introduced by Ferreira (2006). In his book, he described the history and development of this method, explained its theory and presented in detail how to use it.

Organizations matrix

Because the Organizations matrix (OM) is the smallest matrix among the six matrices, the size of its SOM display is the smallest, too. The SOM display of *OM* is demonstrated in Figure 4, and the articles in each cluster are displayed in Table VI. There is a large cluster, three small clusters each including less than four articles and one isolated article in this category. C1 and C2 are close to each other, and both located in the blue area. This indicates that the articles in these two clusters are similar to each other. C4, located in a red area, means it differs from the other clusters. The articles in C1, C2 and C3 are mainly related to surveillance and security projects (e.g. COINTELPRO, DARPA and National Security Agency). These projects aim to promote national security by monitoring behaviour and activities (e.g. NSA warrantless surveillance), collecting and analysing information (e.g. Stellar Wind) or developing new technologies (e.g. Heterogeneous Aerial Reconnaissance Team). The three articles in C4 relate to online information surveillance, which is a relatively

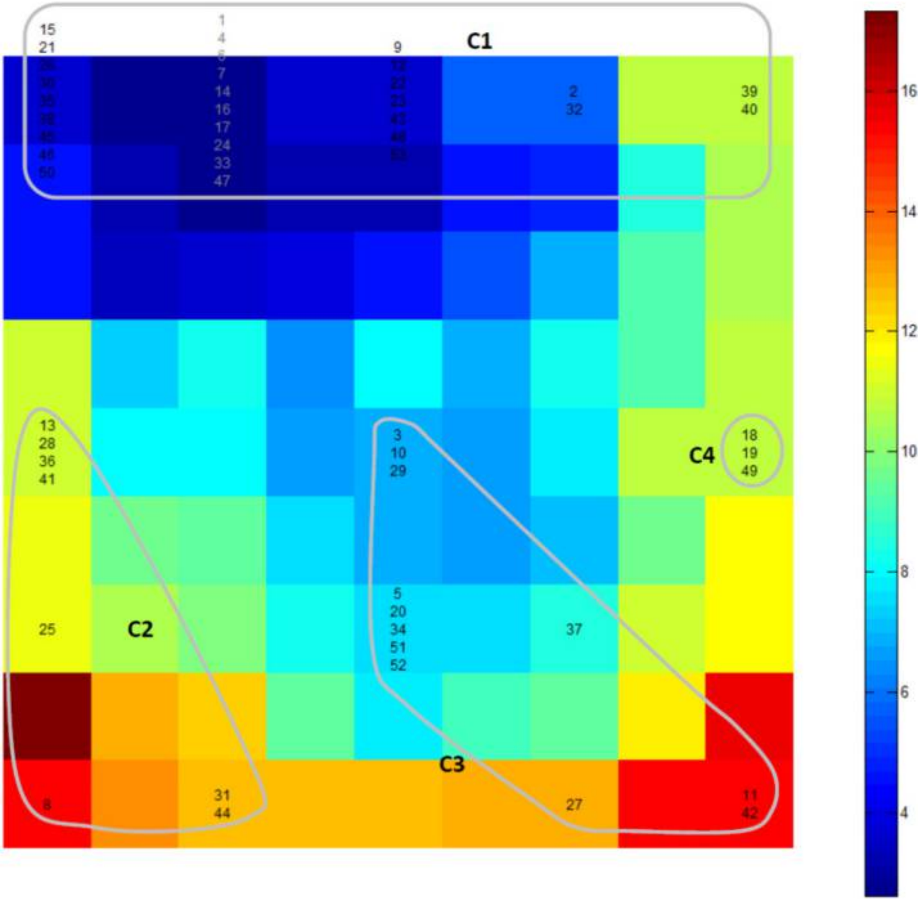

Figure 5.
The Related Fields and Topics matrix display

new area compared to the other clusters. Differing from the projects mentioned before, the Big Brother Watch and the Electronic Frontier Foundation found in C3 campaign against surveillance and for protecting civil liberties. Some other organizations are software-developing companies (e.g. Mattersight Corporation and OpenText).

Related Fields & Topics matrix

The *RFTM* contains the term frequencies of the articles associated with the data-mining-related fields and topics. Its SOM display is presented in Figure 5 and the articles in each cluster are displayed in Table VII. There are three large clusters and one small cluster in this category. Because C2 and C4 are in the green and red areas, respectively, they are quite different from each other.

After investigating the articles of C1, it shows that they are associated to neural network (e.g. ADALINE), biology (e.g. biocybernetics), data analysis (e.g. exploratory data analysis) and security (e.g. phone surveillance). Regarding C2, the articles are related to privacy (e.g. information privacy) and security (e.g. surveillance art). For the articles of C3, artificial intelligence and cognitive science are the primary topics. Meanwhile, the presence of traffic

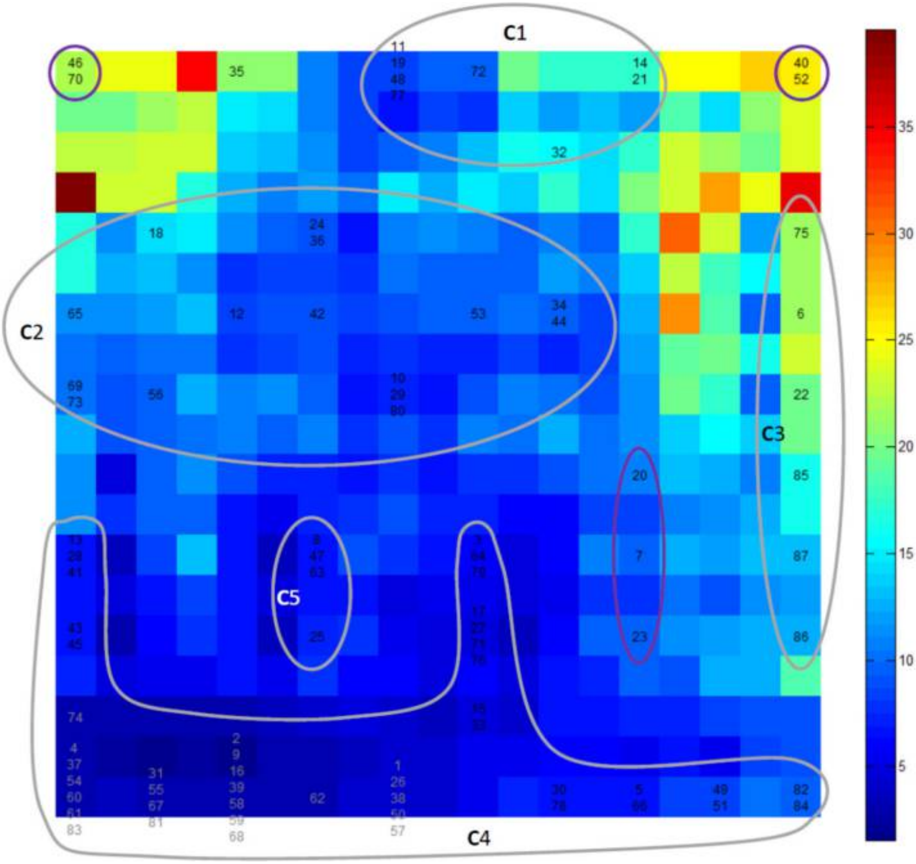

Figure 6.
The Technology Supporting matrix display

analysis indicates that more and more areas cast their eyes on data mining. Among the items obtained, forensic profiling is a unique concept referring to forensic science and security. It is the way to utilize data-mining techniques to explore the patterns of illegal activities (Rannenberg *et al.*, 2009).

Technology Supporting matrix

The TSM contains the term frequencies of the articles associated with technology supporting data mining. The SOM display of *TSM* is presented in Figure 6, and the items in each cluster are displayed in Table VIII. There are four large clusters in this figure, while the isolated numbers and small clusters are grouped into the "Other" group.

All the clusters include articles relevant to software (e.g. neural network software, and Magic Lantern) and systems (e.g. management information system and enterprise planning systems). C1 contains articles about data architecture (e.g. data presentation architecture),= while items of C3 are related to databases (e.g. database model) and languages (e.g. SQL). Apart from software, systems, databases and languages, C4 contains many items that refer to data analysis tools (e.g. Decision Tree and Star Schema).

Cluster	Article
C1	(7) Business analytics, (9) Business Intelligence 2.0, (10) Business semantics management, (14) Computational biology, (15) Computational genomics, (20) Data mining in meteorology, (22) Decision support system, (28) Enterprise search, (30) Functional genomics, (34) List of genetic algorithm applications, (35) Location intelligence, (43) Modelling biological systems, (67) System identification
C2	(11) Category applied data mining, (12) Cheminformatics, (16) Criminal reduction utilizing statistical history, (17) Customer analytics, (18) Customer data integration, (19) Data mining in agriculture, (21) Data warehouse appliance, (23) Demographics, (26) Early case assessment, (40) Master data management, (41) Meteorological intelligence, (47) Online video analytics, (48) Onomastics, (49) Ontology-based data integration, (57) Product management, (59) Psychographics, (63) Runtime intelligence, (64) Sales intelligence, (68) Topological combinatorics
C3	(1) Analytics, (2) Applications of artificial intelligence, (3) Artificial intelligence, (4) Big data, (5) Bioinformatics, (6) Biomedical informatics, (8) Business intelligence, (24) Design of experiments, (25) Drug discovery, (27) Educational data mining, (33) Learning analytics, (38) Mass surveillance in the United Kingdom, (44) Nearest neighbour search, (45) Neural networks, (46) News analytics, (50) Operational intelligence, (52) Pen register, (54) Police-enforced ANPR in the UK, (55) Predictive analytics, (56) Prescriptive analytics, (58) Proteomics, (60) Qualitative research, (61) Real-time business intelligence, (62) Recommendation system, (65) Signals intelligence
Other	(13) Closed-circuit television (CCTV), (31) Information retrieval, (42) Mobile business intelligence, (66) Surveillance, [(29) Expert system, (32) Knowledge discovery], [(36) Machine learning, (53) Phylogenetics], [(37) Mass surveillance, (39) Mass surveillance in the United States], [(51) Operations research, (69) Web analytics]

Table III.
The clusters of the Applications matrix

Cluster	Article
C1	(5) Class membership probabilities, (9) Core data integration, (13) Data curation, (15) Data element, (16) Data fusion, (18) Data mart, (20) Data transformation, (21) Data virtualization, (22) Data wrangling, (26) Digital morphogenesis, (27) Digital traces, (29) Edge data integration, (31) Entity linking, (32) Explanation facility, (36) Identification (information), (40) Information integration, (42) Intelligent document, (43) Intention mining, (44) Labelling, (47) Model transformation, (51) Nothing to hide argument, (52) Online deliberation, (55) Profiling, (57) Referential integrity, (58) Refinement (computing), (60) Security culture, (61) Semantic integration, (62) Semantic translation, (64) Sequence clustering, (65) Sequence labelling, (72) Stop words, (75) Unstructured data, (76) User profile, (77) W-shingling
C2	(2) Analysis, (8) Computational science, (10) Curse of dimensionality, (11) Data acquisition, (19) Data migration, (24) Dataspaces, (30) Enterprise application integration, (39) Information extraction, (41) Integration Competency Center
C3	(4) Buyer decision processes, (34) Fuzzy logic, (53) Pattern recognition, (67) Slowly changing dimension, (70) Statistical classification, (73) String (computer science), (74) Time series
C4	(1) AI effect, (3) Anomaly detection, (17) Data mapping, (63) Sensor fusion, (68) Smart tag (Microsoft)
C5	(7) Clustering high-dimensional data, (49) Named-entity recognition, (50) Noisy text analytics, (59) Schema matching, (66) Sequential pattern mining
Other	(33) Extract, transform, load, (38) Identity transform, (54) Privacy, (56) Radio-frequency identification (RFID), (69) Sousveillance, [(6) Classification rule, (28) Dimension reduction], [(12) Data conversion, (48) Name resolution], [(14) Data deduplication, (25) Digital identity, (35) Habituation], [(23) Database management system, (45) Memristor, (46) Metadata], [(37) Identity (social science), (71) Stereotype]

Table IV.
The clusters of Foundation & Concepts matrix

Cluster	Article
C1	(1) Anscombe's quartet, (2) Approximate nonnegative matrix factorization, (6) BioCreative, (7) Category data clustering algorithms, (9) Change detection, (11) Cluster-weighted modelling, (16) Consensus clustering, (18) CP decomposition, (19) Curve fitting, (24) Forecasting, (27) Fraction of variance unexplained, (30) Group method of data handling, (31) Higher-order factor analysis, (32) Higher-order singular value decomposition, (33) Idea networking, (34) In situ adaptive tabulation, (35) Independent component analysis, (37) JXTA, (38) Kernel machines, (39) Kriging, (41) Local regression, (44) Modifiable areal unit problem, (45) Morphological analysis, (46) Multidimensional scaling, (47) Multilinear algebra, (48) Multilinear PCA, (49) Multilinear subspace learning, (51) Multivariate normal distribution, (52) Multiway data analysis, (53) Neighbourhood components analysis, (54) Network theory, (56) Non-negative matrix factorization, (57) Odds algorithm, (59) Outliers in statistics, (61) Parallel coordinates, (62) Pearson product-moment correlation coefficient, (67) Q methodology, (68) Regression analysis, (69) Regularization perspectives on support vector machines, (70) Robust regression, (74) Sequential minimal optimization, (75) Silhouette (clustering), (76) Spectral clustering, (78) Stepwise regression, (79) Structured data analysis (statistics), (82) Test method, (83) Trend estimation, (84) Tucker decomposition, (86) Varimax rotation, (87) Wavelet, (88) Winnow (algorithm)
C2	(13) Conceptual clustering, (14) Configural frequency analysis, (20) Data stream clustering, (36) Instance-based learning, (40) Latent class analysis, (71) Segmented regression, (72) Semantic warehousing, (80) Systolic array
C3	(3) Association rule learning, (50) Multivariate adaptive regression splines, (55) Nonlinear system identification, (64) Predictive modelling
C4	(23) Factor analysis, (42) Markov chain, (65) Principal component analysis, (77) Statistics, (81) Tensor
C5	(4) Backpropagation, (10) Cluster analysis, (15) Connectionism, (21) Data vault nodelling, (25) Formal concept analysis, (58) Online analytical processing, (66) Process modelling
Other	(8) Censoring (statistics), (26) Fourier analysis, (85) Universal Darwinism, [(5) Binary classification, (43) Metaheuristics, (73) Sequence analysis (Bioinformatics)], [(12) Complex event processing, (17) Constrained clustering, (22) Decision theory], [(28) Gene expression programming, (29) Genetic algorithm], [(60) Paired difference test, (63) Prediction interval]

Table V.
The clusters of Methodologies matrix

Cluster	Article
C1	(4) COINTELPRO, (9) Geodi, (14) List of government surveillance projects
C2	(1) Big Brother Watch, (2) Bioinformatics companies/List of bioinformatics companies, (5) DARPA, (6) DARPA TIPSTER Program, (7) Dropmire, (8) Electronic Frontier Foundation, (10) Government Communications Security Bureau, (11) Heterogeneous Aerial Reconnaissance Team, (12) Integration Consortium, (13) International Society for Computational Biology, (15) Mattersight Corporation, (18) NSA warrantless surveillance (2001-07)/NSA warrantless surveillance controversy, (20) Open text/Open Text Corporation, (21) Operation Ivy Bells, (22) President's Surveillance Program, (25) Stellar Wind, (26) Terrorist Surveillance Program/NSA electronic surveillance program, (27) Total Information Awareness, (28) Trapwire
C3	(16) National Security Agency, (24) Special Communications Service of Russia
C4	(17) NSA in popular culture, (19) Nutch, (23) PRISM
Other	(3) Bullrun (code name)

Table VI.
The clusters of Organizations matrix

Analysis of the themes and topics

As mentioned before, the six categories obtained stand for the six aspects of data mining found in a Wikipedia search. Among the 402 data-mining-related entries, a majority of the entries are about applications, theory foundation and concepts and methodologies.

Table VII.
The clusters of Related Fields and Topics matrix

Cluster	Article
C1	(1) ADALINE, (2) Adaptive resonance theory, (4) Autoencoder, (6) Biocybernetics, (7) Biologically inspired computing, (9) Cerebellar model articulation controller, (12) Computational physics, (14) Concept mining, (15) Connectionist expert system, (16) Connectomics, (17) Customer dynamics, (21) Data retention, (22) Document classification, (23) Exploratory data analysis, (24) Forensic profiling, (26) Informational self-determination, (30) List of scientific journals in bioinformatics, (32) Neuroevolution, (33) Optical neural network, (35) Phone surveillance, (38) Propagation of schema, (39) Radial basis function network, (40) Recurrent neural networks, (43) Spend management, (45) Surveillance system monitor, (46) Telephone tapping in the Eastern Bloc, (47) Tensor product network, (48) Test and learn, (50) Time delay neural network, (53) Web service
C2	(8) Broken windows theory, (13) Computer and network surveillance, (25) Information privacy, (28) Lawful interception, (31) National security, (36) Privacy law, (41) Right to privacy, (44) Surveillance art
C3	(3) Artificial life, (5) BEAM robotics, (10) Cognitive architecture, (11) Cognitive science, (20) Data integration, (27) Judge-advisor system, (29) List of free online bioinformatics courses, (34) Outline of artificial intelligence, (37) Profiling (information science), (42) Sequence alignment, (51) Traffic analysis, (52) Web mining
C4	(18) Data analysis, (19) Data governance, (49) Text mining

Table VIII.
The clusters of the Technology Supporting matrix

Cluster	Article
C1	(11) Business process management, (14) Clinical decision support system, (19) Data presentation architecture, (21) Data warehousing, (32) Enterprise architecture framework, (48) List of software engineering topics, (72) Software as a service, (77) Support vector machines
C2	(10) Business process discovery, (12) Business service management, (18) Cultured neuronal networks, (24) Decision engineering, (29) Digital signal processing, (34) Enterprise integration, (36) Executive information system, (42) Information server, (44) Intrusion detection system, (53) Management information system, (56) Neural network software, (65) Quantitative structure-activity relationship, (69) Self-organizing map, (73) Software for molecular mechanics modelling, (80) Tracking system
C3	(6) AWK (tabular data transforms), (22) Database model, (75) SQL, (85) Web scraping, (86) XQuery, (87) XSLT
C4	(1) Accounting intelligence, (2) Analytic applications, (3) Anchor modelling, (4) Artificial intelligence marketing, (5) ATLAS Transformation Language, (9) Business intelligence tools, (13) Carnivore (software), (15) Compound term processing, (16) Computer and Internet protocol address verifier, (17) Content-addressable memory, (26) Decision tables, (27) Decision tree, (28) Decision tree model, (30) DRAKON, (31) Encog, (33) Enterprise information integration, (37) Expectiminimax tree, (38) Faceted search, (39) Fisher kernel, (41) Influence diagram, (43) Integrated business planning, (45) Land Allocation Decision Support System, (49) Magic Lantern (software), (50) Mail cover, (51) Mail Isolation Control and Tracking, (54) Molecular design software, (55) National Intelligence Priorities Framework, (57) Object-relational mapping, (58) Online transaction processing, (59) Operational data store, (60) Parallel Constraint Satisfaction Processes, (61) Perceptual mapping, (62) Polynomial kernel, (64) Production system, (66) QVT, (67) Relevance vector machine, (68) Self-service software, (71) Snowflake schema, (74) Spatial Decision Support System, (76) Star schema, (78) Tensor software, (79) Three schema approach, (81) Transformation language, (82) Truth table, (83) TXL (programming language), (84) Universal Data Element Framework
C5	(8) Business activity monitoring, (25) Decision making software, (47) List of open-source bioinformatics software, (63) Process mining
Other	(35) Enterprise planning systems, [(7) Behavioural targeting, (20) Data scraping, (23) Datalog], [(40) Government databases, (52) MAINWAY], [(46) List of emerging technologies, (70) Smart grid]

There are multiple popular topics in each aspect. Several strong topics can be found in more than one category: surveillance and security occur in four categories including the Applications, Foundation & Concepts, Organizations and Related Fields & Topics categories; a biological theme occurs in both Applications and in Related Fields & Topics categories; business emerges in the Applications and Technology Supporting categories; artificial intelligence appears in the Applications, Methodologies and Related Fields & Topics categories. The findings reveal that Wikipedia editors and users pay more attention to these data-mining-related themes and topics.

General public versus researchers

As it was stated before, the Wikipedia content reflects the general public's interests. There are similarities and differences between the general public and researchers' interests based on the results obtained. Similar to the Wikipedia content, topics about business and biology have close relationships with data mining in academic research. For examples, business intelligence and biometrics are both important research areas which developed quickly (Chen *et al.*, 2012; Dunstone and Yager, 2008).

In the Applications category, there are two unique topics: meteorology and education. These applications can also be observed in research studies. For example, in meteorology, data mining has been applied to predict storms, forest fires, power of wind farms and so forth (Cortez and de Morais, 2007; Kusiak *et al.*, 2009); educational data mining analysed the data collected from an educational setting by data-mining methods.

Based on the entries in the Methodologies category, it is found that most of the data-mining methods stem from mathematics, statistics, artificial intelligence and neural network. Similar statements can be found in Han and Kamber's (2006) book. In addition to these fields, they claimed that database systems and information retrieval are the other two areas which generate data-mining methods. Biology is another field where data-mining methods emerge in terms of the Wikipedia data. However, it is not regarded as an important field in the previous research studies.

Regarding the technologies supporting data mining, Liao *et al.* (2012) investigated the data-mining-related articles from 2000 to 2011, and classified the data-mining techniques into nine groups, including neural networks, algorithm architecture, dynamic prediction-based approaches, modelling approaches, intelligence agent systems and so on. This result is slightly different from the clustering results of the Technology Supporting category. One of the reasons is that these researchers did not entirely separate data-mining techniques from data-mining methods. Therefore, these nine categories include both methods and techniques.

The findings obtained from the Organizations category imply that the general public is interested in surveillance projects and software companies. It also suggests that the general public pays attention to campaigns against surveillance. Research papers, however, barely focus on or study those organizations.

The findings show that the general public and researchers have similar interests about data mining to some extent. Although Wikipedia articles are usually regarded as being created by the general public, it is possible that the editors of certain entries are experts, or at least possess some knowledge about the corresponding entries. An instance is that the gene expression programming entry on Wikipedia cites seven works of Ferreira for further interpretation of this method. This conclusion confirms Bruns's (2008) statement that both experts and lay people create content on Wikipedia. Meanwhile, differences can be found between the general public and researchers according to the previous paragraphs. In addition, there are special topics discovered from the Wikipedia articles, such as onomastics, cognitive science and forensic science. In the previous information science papers, the connections between these topics and data mining are rarely discussed. Moreover, data-mining organizations and projects are attractive themes to the general public, but few research papers focus on them.

Applying the SOM approach

The SOM approach has been used in various disciplines, such as finance, medical science, library and information science and so on (Mohamed and Abdelsamea, 2014; Sarlin *et al.*, 2012). This study reveals that to achieve explainable outputs, the sizes of outputs are determined by the numbers of items in the corresponding matrices. When the number is larger, the size of the SOM display needs to be larger. For example, the *OM* contains fewer items than the *MM*, so its SOM display size is much smaller than the *MM*'s. If the size of the SOM display is too large or too small, it will be difficult to cluster items into groups.

Implications of this study

The findings of this study explore the themes and topics about data mining from the general public's perspective. Data mining developed rapidly during the past few decades and it is applied in a variety of research fields. The previous reviews usually talk about the theories, techniques and methods, applications and trends of data mining from the researchers' perspective. This study revealed the general public's interests in data mining, including applications, concepts and theories, methods and technologies, organizations and related topics. Popular concepts, themes, topics and the connections among them were examined. It displays the nature of data mining and can contribute to develop ontology related to data mining. These results demonstrate the similarities and differences between the general public and data-mining researchers. Findings will help researchers gain insight into the general public's perceptions of data mining and discover the gap between the general public and themselves.

From the practical perspective, researchers can build ontology systems based on the data-mining-related ontologies obtained. The associated concepts, themes and topics obtained could be recognized as related terms by the recommendation system. In query searching, these items could be provided by information retrieval systems to help users modify their search queries. The whole picture of the structure for the data-mining-related concepts provides a way for information organization from the general public's view.

Another aspect of implications lies in the research methodology. This study uses the SOM approach to cluster Wikipedia entries, which is a new attempt in the information science field. The experience of using SOM in a mixed research method study will benefit future research studies. Second, researchers will be able to find new techniques and methods, and distinctive topics from these Wikipedia entries. It will potentially provide new research topics and data-mining methods for researchers.

Conclusion

Data mining has been a fast-growing research area in recent years. With the incredible increase in the volume of data, data-mining methods are used in more and more disciplines, not only in research but also in people's daily lives. Therefore, both researchers and the general public pay attention to data mining and its related topics.

This study examined the articles relevant to data mining on Wikipedia, and concluded that there are six prominent themes of data mining which gain the general public's attention. Data-mining applications, theories and concepts and methodologies are the most popular themes. The Applications category contained three main topics (business, biology and security) and some special topics, such as meteorology, onomastics and education.

Data-mining methods mainly stemmed from mathematics, statistics, artificial intelligence, neural network and biology. Some of the data-mining theories also came from other fields, such as biology, security and privacy, artificial intelligence and cognitive science. The technologies supporting data mining were software, systems, databases, programming languages and so forth. There were 28 investigated entries associated with data-mining organizations and projects. Some were surveillance and security projects, while at least two other projects campaigned against surveillance.

The general public and researchers both have interests in the applications of data mining, the data-mining theories and the data-mining methods and technologies. The general public, however, focuses more on applications in business than in other fields and this differs from the researchers. The general public also shows interests in data-mining organizations and projects, while the research papers seldom study them in-depth.

Because this paper explores topics about data mining on Wikipedia, all the data were collected from Wikipedia. The limitation is that only 402 articles were obtained in terms of their relevance to data mining. It is possible that some related topics and themes are omitted. To overcome these limitations, larger data sets will be collected and analysed in future studies. Moreover, to reveal the general public's focuses of data mining, collecting more data from other social media platforms is necessary.

References

Ahlqvist, T., Bäck, A., Halonen, M. and Heinonen, S. (2008), "Social media road maps exploring the futures triggered by social media", *VTT Tiedotteita-Valtion Teknillinen Tutkimuskeskus*, Vol. 2454, p. 13.

Aljumah, A.A., Ahamad, M.G. and Siddiqui, M.K. (2013), "Application of data mining: diabetes health care in young and old patients", *Journal of King Saud University-Computer and Information Sciences*, Vol. 25 No. 2, pp. 127-136.

An, L., Zhang, J. and Yu, C. (2011), "The visual subject analysis of library and information science journals with self-organizing map", *Knowledge Organization*, Vol. 38 No. 4, pp. 299-320.

Asadi, S., Ghafghazi, S. and Jamali, H. (2013), "Motivating and discouraging factors for Wikipedians: the case study of Persian Wikipedia", *Library Review*, Vol. 62 Nos 4/5, pp. 237-252.

Benkler, Y. and Nissenbaum, H. (2006), "Commons-based peer production and virtue", *Journal of Political Philosophy*, Vol. 14 No. 4, pp. 394-419.

Bernstein, M.S., Bakshy, E., Burke, M. and Karrer, B. (2013), "Quantifying the invisible audience in social networks", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, ACM, New York, NY*, pp. 21-30.

Borgatti, S., Everett, M. and Freeman, L. (2002), *Ucinet for Windows: Software for Social Network Analysis*, Analytic Technologies, Harvard, MA.

Borghini, A., Crotti, P., Pietra, D., Favero, L. and Bianucci, A.M. (2010), "Chemical reactivity predictions: use of data mining techniques for analyzing regioselective azidolysis of epoxides", *Journal of Computational Chemistry*, Vol. 31 No. 14, pp. 2612-2619.

Brabham, D.C. (2008), "Crowdsourcing as a model for problem solving: an introduction and cases", *Convergence*, Vol. 14 No. 1, pp. 75-90.

Brum, J.R., Schenck, R.O. and Sullivan, M.B. (2013), "Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses", *The ISME Journal*, Vol. 7 No. 9, pp. 1738-1751.

Bruns, A. (2006), "Towards produsage: futures for user-led content production", in Sudweeks, F., Hrachovec, H. and Ess, C. (Eds), *Creative Industries Faculty*, presented at the Cultural Attitudes towards Communication and Technology, Murdoch University, Tartu, pp. 275-284.

Bruns, A. (2007), "Produsage", in *Proceedings of the 6th ACM SIGCHI Conference on Creativity & Cognition, C&C '07*, ACM, New York, NY, pp. 99-106.

Bruns, A. (2008), *Blogs, Wikipedia, Second Life, and Beyond: From Production to Produsage*, Peter Lang, New York, NY.

Bryant, S.L., Forte, A. and Bruckman, A. (2005), "Becoming wikipedian: transformation of participation in a collaborative online encyclopedia", in *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work, GROUP '05, ACM, New York, NY*, pp. 1-10.

Callahan, E.S. and Herring, S.C. (2011), "Cultural bias in Wikipedia content on famous persons", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 10, pp. 1899-1915.

Chen, H., Chiang, R.H. and Storey, V.C. (2012), "Business intelligence and analytics: from Big Data to big impact", *MIS Quarterly*, Vol. 36 No. 4, pp. 1165-1188.

Cortez, P. and de Morais, A.J.R. (2007), "A data mining approach to predict forest fires using meteorological data", in Neves, J.M., Santos, M.F. and Machado, J.M. (Eds), *New Trends in Artificial Intelligence: Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA 2007), Guimarães, Associação Portuguesa para a Inteligência Artificial, Lisboa*, pp. 512-523.

Deboeck, G. and Kohonen, T. (2013), *Visual Explorations in Finance: with Self-Organizing Maps*, Springer Science & Business Media, New York, NY.

Doan, A., Ramakrishnan, R. and Halevy, A.Y. (2011), "Crowdsourcing systems on the world-wide web", *Communications of the ACM*, Vol. 54 No. 4, pp. 86-96.

Dunstone, T. and Yager, N. (2008), *Biometric System and Data Analysis: Design, Evaluation, and Data Mining*, Springer Science & Business Media, New York, NY.

Elkan, C. (2001), "Magical thinking in data mining: lessons from CoIL challenge 2000", *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY*, pp. 426-431.

Ferreira, C. (2006), *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*, Springer Verlag, Berlin.

Flanagin, A.J. and Metzger, M.J. (2011), "From Encyclopaedia Britannica to wikipedia: generational differences in the perceived credibility of online encyclopedia information", *Information, Communication & Society*, Vol. 14 No. 3, pp. 355-374.

Fu, T. (2011), "A review on time series data mining", *Engineering Applications of Artificial Intelligence*, Vol. 24 No. 1, pp. 164-181.

Fuertes, J.J., Domínguez, M., Reguera, P., Prada, M.A., Díaz, I. and Cuadrado, A.A. (2010), "Visual dynamic model based on self-organizing maps for supervision and fault detection in industrial processes", *Engineering Applications of Artificial Intelligence*, Vol. 23 No. 1, pp. 8-17.

Han, J. and Kamber, M. (2006), *Data Mining: Concepts and Techniques*, 2nd ed., Morgan Kaufmann, Burlington, MA.

Hansen, D., Shneiderman, B. and Smith, M.A. (2010), *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*, Morgan Kaufmann, Burlington, MA.

Hill, C.A., Dean, E. and Murphy, J. (2013), *Social Media, Sociality, and Survey Research*, John Wiley & Sons, Hoboken, NJ.

Himberg, J., Alhoniemi, E. and Parhankangas, J. (2000), *SOM Toolbox for MATLAB 5*, Helsinki University of Technology, Helsinki.

Howe, J. (2006), "The rise of crowdsourcing", *Wired*, Vol. 14 No. 6, pp. 1-4.

Hu, M., Lim, E.P., Sun, A., Lauw, H.W. and Vuong, B.-Q. (2007), "Measuring article quality in wikipedia: models and evaluation", in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07, ACM, New York, NY*, pp. 243-252.

Iba, T., Nemoto, K., Peters, B. and Gloor, P.A. (2010), "Analyzing the creative editing behavior of Wikipedia editors", *Procedia - Social & Behavioral Science, The 1st Collaborative Innovation Networks Conference - COINs 2009*, *Savannah, Georgia*, Vol. 2, pp. 6441-6456.

Johansson, P., Ekstrand-Tobin, A. and Bok, G. (2014), "An innovative test method for evaluating the critical moisture level for mould growth on building materials", *Building and Environment*, Vol. 81, pp. 404-409.

Kaplan, A.M. and Haenlein, M. (2010), "Users of the world, unite! The challenges and opportunities of social media", *Business Horizons*, Vol. 53 No. 1, pp. 59-68.

Kietzmann, J.H., Hermkens, K., McCarthy, I.P. and Silvestre, B.S. (2011), "Social media? Get serious! Understanding the functional building blocks of social media", *Business Horizons*, Vol. 54 No. 3, pp. 241-251, (Special Issue: Social Media).

Kohonen, T. (1990), "The self-organizing map", *Proceedings of the IEEE*, Vol. 78 No. 9, pp. 1464-1480.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V. and Saarela, A. (2000), "Self organization of a massive document collection", *IEEE Transactions on Neural Networks*, Vol. 11 No. 3, pp. 574-585.

Kolaczyk, E.D. and Csárdi, G. (2014), *Statistical Analysis of Network Data with R, Use R!*, Springer, New York, NY.

Kusiak, A., Zheng, H. and Song, Z. (2009), "Short-term prediction of wind farm power: a data mining approach", *IEEE Transactions on Energy Conversion*, Vol. 24 No. 1, pp. 125-136.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010), "What is Twitter, a social network or a news media?", *Proceedings of the 19th International Conference on World Wide Web*, *ACM, New York, NY*, pp. 591-600.

Laniado, D., Kaltenbrunner, A., Castillo, C. and Morell, M.F. (2012), "Emotions and dialogue in a peer-production community: the case of Wikipedia", *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, WikiSym '12*, *ACM, New York, NY*, pp. 1-9:10.

Law, M.H. and Kwok, J.T. (2000), "Rival penalized competitive learning for model-based sequence clustering", *Proceedings of the 15th International Conference on Pattern Recognition*, *Barcelona*, Vol. 2, pp. 195-198.

Liao, S.H., Chu, P.H. and Hsiao, P.Y. (2012), "Data mining techniques and applications: a decade review from 2000 to 2011", *Expert Systems with Applications*, Vol. 39 No. 12, pp. 11303-11311.

Lin, X., Soergel, D. and Marchionini, G. (1991), "A self-organizing semantic map for information retrieval", *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '91*, *ACM, New York, NY*, pp. 262-269.

Miner, G. (2012), *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Academic Press, Cambridge, MA.

Mohamed, M.H. and Abdelsamea, M.M. (2014), "Self organization map based texture feature extraction for efficient medical image categorization", *Proceedings of the 4th ACM International Conference on Intelligent Computing and Information Systems, ICICIS 2009*, *Cairo, ACM, New York, NY*, pp. 943-948.

Moskaliuk, J., Kimmerle, J. and Cress, U. (2012), "Collaborative knowledge building with wikis: the impact of redundancy and polarity", *Computers & Education*, Vol. 58 No. 4, pp. 1049-1057.

de Nooy, W., Mrvar, A. and Batagelj, V. (2011), *Exploratory Social Network Analysis with Pajek*, Cambridge University Press, Cambridge.

Paltoglou, G. (2014), "Sentiment analysis in social media", in Agarwal, N., Lim, M. and Wigand, R.T. (Eds), *Online Collective Action, Lecture Notes in Social Networks*, Springer, Vienna, pp. 3-17.

Pandit, N.R. (1996), "The creation of theory: a recent application of the grounded theory method", *The Qualitative Report*, Vol. 2 No. 4, pp. 1-15.

Rannenberg, K., Royer, D. and Deuker, A. (2009), *The Future of Identity in the Information Society: Challenges and Opportunities*, Springer Science & Business Media, New York, NY.

Sarlin, P., Yao, Z. and Eklund, T. (2012), "A framework for state transitions on the self-organizing map: Some temporal financial applications", *Intelligent Systems in Accounting, Finance and Management*, Vol. 19 No. 3, pp. 189-203.

Spaho, E., Sakamoto, S., Barolli, L., Xhafa, F. and Ikeda, M. (2013), "Trustworthiness in P2P: performance behaviour of two fuzzy-based systems for JXTA-overlay platform", *Soft Computing*, Vol. 18 No. 9, pp. 1783-1793.

Suchanek, F.M., Sozio, M. and Weikum, G. (2009), "SOFIE: a self-organizing framework for information extraction", *Proceedings of the 18th International Conference on World Wide Web, WWW '09, ACM, New York, NY*, pp. 631-640.

Thelwall, M., Buckley, K. and Paltoglou, G. (2011), "Sentiment in Twitter events", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 2, pp. 406-418.

Ultsch, A. and Siemon, H. (1990), "Kohonen's self organizing feature maps for exploratory data analysis", *presented at the Proceedings of INNC'90, International Neural Network Conference, Kluwer*, pp. 305-308.

Viera, A.J. and Garrett, J.M. (2005), "Understanding interobserver agreement: the kappa statistic", *Family Medicine*, Vol. 37 No. 5, pp. 360-363.

Westaby, J.D., Pfaff, D.L. and Redding, N. (2014), "Psychology and social networks: a dynamic network theory perspective", *American Psychologist*, Vol. 69 No. 3, pp. 269-284.

Zhang, J. (2007), *Visualization for Information Retrieval*, Springer Science & Business Media, New York, NY.

Zhang, L., Zhang, L., Teng, W. and Chen, Y. (2013), "Based on information fusion technique with data mining in the application of finance early-warning", *Procedia Computer Science*, Vol. 17, pp. 695-703.

Zhao, S.J., Zhang, K.Z.K., Wagner, C. and Chen, H. (2013), "Investigating the determinants of contribution value in Wikipedia", *International Journal of Information Management*, Vol. 33 No. 1, pp. 83-92.

About the authors

Yanyan Wang is a full-time Doctoral Candidate at the School of Information Studies, University of Wisconsin-Milwaukee, USA. Her research interests include information retrieval, social media, text mining and user behaviour. Yanyan Wang is the corresponding author and can be contacted at: wang238@uwm.edu

Dr Jin Zhang is a Full Professor at the School of Information Studies, University of Wisconsin-Milwaukee, USA. His research interests include visualization for information retrieval, information retrieval theory and algorithm, metadata, search engine evaluation, consumer health informatics, transaction log analysis, digital libraries and computer–human interface design. He has published papers in journals such as *JASIST*, *IPM*, *Journal of Intelligent Information Systems*, *Journal of Information Retrieval*, *Journal of Information Science*, and *IEEE Transactions on Visualization and Computer Graphics*. His book *Visualization for Information Retrieval* was published in the Information Retrieval Series by Springer in 2008.