Chapter 6
Big Data Applications

Abstract In the previous chapter, we examined big data analysis, which is the final and most important phase of the value chain of big data. Big data analysis can provide useful values via judgments, recommendations, supports, or decisions. However, data analysis involves a wide range of applications, which frequently change and are extremely complex. In this chapter, the evolution of data sources is reviewed. Then, six of the most important data analysis fields are examined, including structured data analysis, text analysis, website analysis, multimedia analysis, network analysis, and mobile analysis. This chapter is concluded with a discussion of several key application fields of big data.

6.1 Application Evolution

Recently, big data and big data analysis has been proposed for describing datasets and as analytical technologies in large-scale complex programs, which need to be analyzed with advanced analytical methods. As a matter of fact, data driven applications have emerged in the past decades. For example, as early as 1990s, business intelligence has became a prevailing technology for business applications and, network search engines based on massive data mining processing emerged in the early twenty-first century. Some potential and influential applications from different fields and their data and analysis characteristics are discussed as follows.

- *Evolution of Commercial Applications*: The earliest business data was generally structured data, which was collected by companies from old systems and then stored in RDBMSs. Analytical technologies used in such systems were prevailing in 1990s and was intuitive and simple, e.g., reports, instrument panels, special queries, search-based business intelligence, online transaction processing, interactive visualization, score cards, predictive modeling, and data mining [1]. Since the beginning of twenty-first century, networks and websites has been providing a unique opportunity for organizations to have online display and directly interact

M. Chen et al., *Big Data: Related Technologies, Challenges and Future Prospects*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-06245-7_6,
© The Author(s) 2014

with customers. Abundant products and customer information, including click stream data logs and user behavior, etc., can be acquired from the websites. Product layout optimization, customer trade analysis, product suggestions, and market structure analysis can be conducted by text analysis and website mining technologies. As reported in [2], the quantity of mobile phones and tablet PC first surpassed that of laptops and PCs in 2011. Mobile phones and Internet of Things based on sensors are opening a new generation of innovation applications, and searching for larger capacity of supporting location sensing, people oriented, and context operation.

- *Evolution of Network Applications*: The early Internet mainly provided email and webpage services. Text analysis, data mining, and webpage analysis technologies have been applied to the mining of email contents and building search engines. Nowadays, most applications are web-based, regardless of their application field and design goals. Network data accounts for a major percentage of the global data volume. Web has became a common platform for interconnected pages, full of various kinds of data, such as text, images, videos, pictures, and interactive contents, etc. Therefore, plentiful advanced technologies used for semi-structured or unstructured data emerged at the right moment. For example, the image analysis technology may extract useful information from pictures, e.g., face recognition. Multimedia analysis technologies can be applied to the automated video surveillance systems for business, law enforcement, and military applications. Since 2004, online social media, such as Internet forums, online communities, blogs, social networking services, and social multimedia websites, etc., provide users with great opportunities to create, upload, and share contents generated by users. Different user groups may search for daily news and celebrity news, publish their social and political opinions, and provide different applications with timely feedback.
- *Evolution of Scientific Applications*: Scientific research in many fields is acquiring massive data with high-throughput sensors and instruments, such as astrophysics, oceanology, genomics, and environmental research. The U.S. National Science Foundation (NSF) has recently announced the BIGDATA Research Initiative to promote research efforts to extract knowledge and insights from large and complex collections of digital data. Some scientific research disciplines have developed massive data platforms and obtained useful outcomes. For example, in biology, iPlant [3] applies network infrastructure, physical computing resources, coordination environment, virtual machine resources, and inter-operative analysis software and data service to assist researches, educators, and students in enriching all plant sciences. IPlant dataset have high varieties in form, including specification or reference data, experimental data, analog or model data, observation data, and other derived data.

6.2 Big Data Analysis Fields

Data analysis research can be divided into six key technical fields, i.e., structured data analysis, text data analysis, website data analysis, multimedia data analysis, network data analysis, and mobile data analysis. Such a classification aims to emphasize data characteristics, but some of the fields may utilize similar technologies. Since data analysis has a broad scope and it is not easy to have a comprehensive coverage, we will focus on the key problems and technologies in data analysis in the following discussions.

6.2.1 Structured Data Analysis

Business applications and scientific research may generate massive structured data, of which the management and analysis rely on mature commercialized technologies, such as RDBMS, data warehouse, OLAP, and BPM (Business Process Management) [4]. Data analysis is mainly based on data mining and statistical analysis, both of which have been well studied over the past 30 years.

Data analysis is still a very active research field and new application demands drive the development of new methods. Statistical machine learning based on exact mathematical models and powerful algorithms have been applied to anomaly detection [5] and energy control [6]. Exploiting data characteristics, time and space mining may extract knowledge structures hidden in high-speed data flows and sensor data models and modes [7]. Driven by privacy protection in e-commerce, e-government, and health care applications, privacy protection data mining is an emerging research field [8]. Over the past decade, benefited by the substantial popularization of event data, new process discovery, and consistency check technologies, process mining is becoming a new research field especially in process analysis with event data [9].

6.2.2 Text Data Analysis

The most common format of information storage is text, e.g., email communication, business documents, web pages, and social media. Therefore, text analysis is deemed to feature more business-based potential than structured data mining. Generally, tax analysis, also called text mining, is a process to extract useful information and knowledge from unstructured text. Text mining is an inter-disciplinary problem, involving information retrieval, machine learning, statistics, computing linguistics, and data mining in particular. Most text mining systems are based on text expressions and natural language processing (NLP), with more focus on the latter.

Document introduction and query processing are the foundation for developing vector space model, Boolean Retrieval Model, and probability retrieval model, which constitute the foundation of search engines. Since the early 1990s, search engines have evolved into a mature business system, which generally consist of rapidly distributed crawling, effectively inverted index, webpage sequencing based on inlink, and search log analysis [10].

NLP can enable computers to analyze, interpret, and even generate text. Some common NLP methods are: lexical acquisition, word sense disambiguation, part-of-speech tagging, and probabilistic context free grammar [11]. Some NLP-based technologies have been applied to text mining, including information extraction, topic models, text summarization, classification, clustering, question answering, and opinion mining. Information mining shall automatically extract specific structured information from texts. Named entity recognition (NER) technology, as a subtask of information extraction, aims to recognize atomic entities in texts subordinate to scheduled categories (e.g. figures, places, and organizations), which have been successfully applied to the development of new analysis [12] and medical applications [13] recently. The topic models are built according to the opinion that "documents are constituted by topics and topics are the probability distribution of vocabulary." Topic models are models generated by documents, stipulating the probability program to generate documents.

Presently, various probabilistic topic models have been used to analyze document contents and lexical meanings [14]. Text summarization is to generate a reduced summary or extract from a single or several input text files. Text summarization may be classified into concrete summarization and abstract summarization [15]. Concrete summarization selects important sentences and paragraphs from source documents and concentrates them into shorter forms. Abstract summarization may interpret the source texts and, according to linguistic methods, use a few words and phrases to represent the source texts.

Text classification is to recognize probabilistic topic of documents by putting documents in scheduled topics. Text classification based on the new graph representation and graph mining has recently attracted considerable interest [16]. Text clustering is used to group similar documents with scheduled topics, which is different from text classification that gathers documents together. In text clustering, documents may appear in multiple subtopics. Generally, some clustering algorithms in data mining can be utilized to compute the similarities of documents. However, it is also shown that the structural relationship information may be exploited to improve the clustering performance in Wikipedia [17]. The question answering system is designed to search for the optimal answer to a given question. It involves different technologies of question analysis, source retrieval, answer extraction, and answering demonstration [18]. The question answering system may be applied in many fields, including education, website, healthcare, and national defense. Opinion mining, similar to sentiment analysis, refers to the computing technologies for identifying and extracting subjective information from news assessment, comment,

and other user-generated contents. It provides opportunities for users to understand the opinions of the public and customers on social events, political movements, business strategies, marketing activities, and product preference [19].

6.2.3 Web Data Analysis

Over the past decade, we have witnessed the explosive growth of Internet information. Web analysis has emerged as an active research field. Web analysis aims to automatically retrieve, extract, and evaluate information from Web documents and services so as to discover useful knowledge. Web analysis is related to several research fields, including database, information retrieval, NLP, and text mining. According to the different parts of the Web to be mined, we classify Web analysis into three related fields: Web content mining, Web structure mining, and Web usage mining [20].

Web content mining is the process to discover useful knowledge in Web pages, which generally involve several types of data, such as text, image, audio, video, code, metadata, and hyperlink.

The research on image, audio, and video mining has recently been called multimedia analysis, which will be discussed in Sect. 6.2.4. Since most Web content data is unstructured text data, the research on Web data analysis mainly centers around text and hypertext. Text mining is discussed in Sect. 6.2.2, while Hypertext mining involves mining semi-structured HTML files that contain hyperlinks.

Supervised learning and classification play important roles in hyperlink mining, e.g., email, newsgroup management, and Web catalogue maintenance [21]. Web content mining can be conducted with two methods: the information retrieval method and the database method. Information retrieval mainly assists in or improves information lookup, or filters user information according to deductions or configuration documents. The database method aims to simulate and integrate data in Web, so as to conduct more complex queries than searches based on key words.

Web structure mining involves models for discovering Web link structures. Here, the structure refers to the schematic diagrams linked in a website or among multiple websites. Models are built based on topological structures provided with hyperlinks with or without link description. Such models reveal the similarities and correlations among different websites and are used to classify website pages. Page Rank [22] and CLEVER [23] make full use of the models to look up related website pages. Topic-oriented crawler is another successful case by utilizing the models [24]. Topic-oriented crawler is targeted at selectively discovering pages related to scheduled topic sets. Top-oriented crawler may analyze crawling boundary to look for links mostly related to crawling and to avoid the involvement of irrelevant areas, other than collecting and indexing all accessible webpage files, so as to answer all possible Ad-Hoc queries. This way, a great quantity of hardware and network resources may be saved and crawling updating task may be assisted.

Web usage mining aims to mine auxiliary data generated by Web dialogues or behaviors. Web content mining and Web structure mining use the master Web data. Web usage data includes access logs at Web servers, logs at proxy servers, browsers' history records, user profiles, registration data, user sessions or trades, cache, user queries, bookmark data, mouse click and scroll, and any other kind of data generated through interaction with the Web. As Web services and the Web2.0 are becoming mature and popular, Web usage data will have increasingly high variety. Web usage mining plays key roles in personalized space, e-commerce, network privacy/security, and other emerging fields. For example, collaborative recommender systems can personalize e-commerce by utilizing the different preferences of users.

6.2.4 Multimedia Data Analysis

Multimedia data (mainly including images, audios, and videos) have been growing at an amazing speed. Multimedia content sharing is to extract related knowledge and understand semantemes contained in multimedia data. Because multimedia data is heterogeneous and most of such data contains richer information than simple structured data and text data, extracting information is confronted with the huge challenge of the semantic differences of multimedia data. Research on multimedia analysis covers many disciplines. Some recent research priorities include multimedia summarization, multimedia annotation, multimedia index and retrieval, multimedia suggestion, and multimedia event detection, etc.

Audio summarization can be accomplished by simply extracting the prominent words or phrases from metadata or synthesizing a new representation. Video summarization is to interpret the most important or representative video content sequence, and it can be static or dynamic. Static video summarization methods utilize a key frame sequence or context-sensitive key frames to represent a video. Such methods are very simple and have been applied to many business applications (e.g., Yahoo!, Alta Visa, and Google), but the playback performance is poor. Dynamic summarization methods use a series of video clips to represent a video, configure low-level video functions, and take other smooth measures to make the final summarization look more natural. In [25], the authors proposed a topic-oriented multimedia summarization system (TOMS) that can automatically summarize the important information in a video belonging to a certain topic area, based on a given set of extracted features from the video.

Multimedia annotation inserts labels to describe contents of images and videos in both syntax and semantic levels. With the assistance of such labels, the management, summarization, and retrieval of multimedia data can be easily implemented. Since manual annotation is both time and labor intensive, multimedia automatic annotation without any human interventions becomes highly appealing. The main challenge for multimedia automatic annotation is semantic difference, i.e. the difference between low-level features and annotations. Although much progress has

been made, the performance of the existing automatic annotation methods still needs to be improved. Currently, many efforts are being made to synchronously explore both manual and automatic multimedia annotation [26].

Multimedia index and retrieval involve describing, storing, and organizing multimedia information and assisting users to conveniently and quickly look up multimedia resources [27]. Generally, multimedia index and retrieval include five procedures: structural analysis, feature extraction, data mining, classification and annotation, query and retrieval [28]. Structural analysis aims to segment a video into several semantic structural elements, including lens boundary detection, key frame extraction, and scene segmentation, etc. According to the result of structural analysis, the second procedure is feature extraction, which mainly includes further mining the features of necessary key frames, objects, texts, and movements, which are the foundation of video index and retrieval. Data mining, classification, and annotation are generated to utilize the extracted features to find the modes of video contents and put videos into scheduled categories so as to generate video indexes. Upon receiving a query, the system will use a similarity measurement method to look up a candidate video. The retrieval result optimizes the related feedback.

Multimedia recommendation aims to recommend specific multimedia contents according to users' preferences. It is proven to be an effective approach to provide quality personalized services. Most existing recommendation systems can be classified into content-based systems and collaborative-filtering-based systems. The content-based methods identify users or general features in which the users are interested, and recommend users for other contents with similar features. These methods purely rely on content similarity measurement but most of them are limited by content analysis and excessive specifications. The collaborative-filtering-based methods identify groups with similar interests and recommend contents for group members according to their behaviors [29]. Presently, a mixed method is introduced, which integrates advantages of the aforementioned two types of methods to improve the recommendation quality [30].

The U.S. NIST initiated the TREC Video Retrieval Evaluation detecting the occurrence of an event in video-clips based on Event Kit, which contains some text description related to concepts and video examples [31]. The research on video event detection is still in its infancy. The existing research on event detection mainly focuses on sports or news events, running or abnormal events in monitoring videos, and other similar events with repetitive patterns. In [32], the author proposed a new algorithm on special multimedia event detection using a few positive training examples.

6.2.5 Network Data Analysis

Network analysis evolved from the initial quantitative analysis [33] and sociological network analysis [34] into the emerging online social network analysis in the beginning of twenty-first century. Many prevailing online social networking services

include Twitter, Facebook, and LinkedIn, etc. have been increasingly popular over the years. Such online social networking services generally include massive linked data and content data. The linked data is mainly in the form of graphic structures, describing the communications between two entities. The content data contains text, image, and other network multimedia data. The rich contents of such networks bring about both unprecedented challenges and opportunities to data analysis. In accordance with the data-centered perspective, the existing research on social networking service contexts can be classified into two categories: link-based structural analysis and content-based analysis [35].

The research on link-based structural analysis has always been committed on link prediction, community discovery, social network evolution, and social influence analysis, etc. SNS may be visualized as graphs, in which every vertex corresponds to a user and edges correspond to the correlations among users. Since SNS are dynamic networks, new vertexes and edges are continually added to the graphs. Link prediction is to predict the possibility of future connection between two vertexes. Many technologies can be used for link prediction, e.g., feature-based classification, probabilistic methods, and Linear Algebra. Feature-based classification is to select a group of features for a vertex and utilize the existing link information to generate binary classifiers to predict the future link [36]. Probabilistic methods aim to build models for connection probabilities among vertexes in SNS [37]. Linear Algebra computes the similarity between two vertexes according to the singular similar matrix [38]. A community is represented by a sub-graphic matrix, in which edges connecting vertexes in the sub-graph feature high density, while the edges between two sub-graphs feature much lower density [39].

Many methods against community detection have been proposed and studied, most of which are topology-based target functions relying on the concept of capturing community structure. Du et al. utilized the property of overlapping communities in real life to propose a more effective large-scale SNS community detection method [40]. The research on SNS aims to look for a law and deduction model to interpret network evolution. Some empirical studies found that proximity bias, geographical limitations, and other factors play important roles in SNS evolution [41–43], and some generation methods are proposed to assist network and system design [44].

Social influence refers to the case when individuals change their behavior under the influence of others. The strength of social influence depends on the relation among individuals, network distances, time effect, and characteristics of networks and individuals, etc. Marketing, advertisement, recommendation, and other applications can benefit from social influence by qualitatively and quantitatively measuring the influence of individuals on others [45, 46]. Generally, if the proliferation of contents between SNS are considered, the performance of link-based structural analysis may be further improved.

Benefited by the revolutionary progress of Web2.0, the use of generated contents is explosively growing in SNS. SNS is used to generated contents by various technology, including blogs, micro blogs, opinion mining, photos, video sharing, social bookmarking, social network sites, social news, and Wiki. Content-based

analysis in SNS is also known as social media analysis. Social media include text, multimedia, positioning, and comments. Nearly all research topics related to structural analysis, text analysis, and multimedia analysis may be interpreted as social media analysis, but social media analysis is confronted with unprecedented challenges. First, massive and continually growing social media data should be automatically analyzed within a reasonable time. Second, social media data contains much noise, e.g., blogosphere contains a large number of spam blogs, and so does trivial Tweets in Twitter. Third, SNS are dynamic networks, which are frequently and quickly changed and updated.

Since social media is close to SNS, social media analysis is inevitably influenced by SNS analysis. SNS analysis refers to the text analysis of SNS context and characteristics of social and network structures, as well as multimedia analysis. The existing research on social media analysis is still in its infancy. The applications of SNS text analysis include transfer learning in keyword search, classification, clustering, and heterogeneous networks. Keyword search tries to synchronously use contents and link behaviors for search [47]. The motivation for such applications is that text files containing similar keywords are generally connected to each other [48]. During classification, assuming all nodes of the SNS are provided with labels, the nodes added with labels are classified. During clustering, researchers aim to determine node sets with similar contents and accordingly group them [49]. Considering that SNS contains massive information of different interlinked objects, e.g., articles, labels, images, and videos, transfer learning in heterogeneous networks aims to transfer knowledge information among different links [50].

Multimedia datasets in SNS is organized in a structured form, which brings rich information, e.g., semantic ontology, social interaction, community media, geographical maps, and multimedia opinions. Structural multimedia analysis in SNS is also called multimedia information networks. The link structure of multimedia information networks is mainly a logic structure, which are of vital importance to the multimedia in multimedia networks. The logic connection structures in multimedia information networks can be classified into four types: semantic ontology, community media, individual photo albums, and geographical positions [36].

6.2.6 Mobile Traffic Analysis

With the rapid growth of mobile computing, mobile terminals and applications in the world are growing rapidly. By April 2013, Android Apps has provided more than 650,000 applications, covering nearly all categories. By the end of 2012, the monthly mobile data flow has reached 885 PB [51]. The massive data and abundant applications exploit a broad research field for mobile analysis but also bring about a few challenges. As a whole, mobile data has unique characteristics, e.g., mobile sensing, moving flexibility, noise, and a large amount of redundancy. Recently, new research on mobile analysis has been started in different fields. Because of the far

immaturity of the research on mobile analysis, we will only introduce some recent and representative analysis applications in this section.

With the growth of numbers of mobile users and the improved performance, mobile phones are now useful for building and maintaining communities, such as communities based on geographical locations and communities based on different cultures and interests, e.g., the latest Wechat. Traditional network communities or SNS communities are in short of online interaction among members, and the communities are active only when members are sitting before computers. On the contrary, mobile phones can support rich interaction any time and anywhere. Wechat supports not only one-to-one communications, but also many-to-many communication. Mobile communities are defined as that a group of individuals with the same hobbies (i.e., health, safety, and entertainment, etc.) gather together on networks, meet to make a common goal, decide measures through consultation to achieve the goal, and start to implement their plan [52]. In [53], the authors proposed a qualitative model of a mobile community. It is now widely believed that mobile community applications will greatly promote the development of the mobile industry.

RFID labels are used to identify, locate, track, and supervise physical objects in a cost-effective manner. RFID is widely applied to inventory management and logistics. However, RFID brings about many challenges to data analysis: (a) RFID data is very noisy and redundant; (b) RFID data is instant and streaming data with a huge volume and limited processing time. We can track objects and monitor system status by deducing some original events through mining the semantics of RFID data, including location, cluster, and time, etc. In addition, we may design the application logic as complex events and then detect such complex events, so as to realize more advanced business applications. In [54], the authors discussed a shoplifting case as an advanced complex event.

Recently, the progress in wireless sensor, mobile communication technology, and stream processing enable people to build a body area network to have real-time monitoring of people's health. Generally, medical data from different sensors has different characteristics, e.g., heterogeneous attribute sets, different time and space relations, and different physiological relations, etc. In addition, such datasets involve privacy and safety protection. In [55], Garg and others introduced a multi-modal transport analysis mechanism of raw data for real-time monitoring of health. Under the circumstance that only highly comprehensive characteristics related to health are available, Park et al. in [56] examined approaches to better utilize such comprehensive information to strength data at all levels. Comprehensive statistics of some partitions is used to recognize clustering and input a characteristic value with a more comprehensive degree. The input characteristics will be further used to predict modeling so as to improve performance.

Researchers from Gjovik University College in Norway and Derawi Biometrics united to develop an application for smart phones, which analyzes paces when people walk and uses the paces for unlocking the safety system [57]. In the meanwhile, Robert Delano and Brian Parise from Georgia Institute of Technology developed an application called iTrem, which monitors human bodies' trembling

with a built-in seismograph in a mobile phone, so as to cope with Parkinson and other nervous system diseases [57]. Many other mobile device applications aim to acquire information through mobile devices, no matter how useful such information is for future data analysis.

6.3 Key Applications

6.3.1 Application of Big Data in Enterprises

At present, big data mainly comes from and used in enterprises, while BI and OLAP can be regarded as the predecessors of big data application. The application of big data in enterprises can enhance their production efficiency and competitiveness in many aspects. In particular, on marketing, with correlation analysis of big data, enterprises can more accurately predict the behavior of consumers and mine new business modes. On sales planning, after comparison of massive data, enterprises can optimize their commodity prices. On operation, enterprises can improve their operation efficiency and operation satisfaction, optimize the input of labor force, accurately forecast personnel allocation requirements, avoid excess production capacity, and reduce labor cost. On supply chain, using big data, enterprises may conduct inventory optimization, logistic optimization, and supplier coordination, etc., to mitigate the gap between supply and demand, control budgets, and improve services.

In finance, the application of big data in enterprises has been rapidly developed. For example, China Merchants Bank (CMB) utilizes data analysis to recognize that such activities as "Multi-times score accumulation" and "score exchange in shops," are effective for attracting quality customers. By building a customer loss early warning model, the bank can sell high-yield financial products to the top 20 % customers in loss ratio so as to retain them. As a result, the loss ratios of customers with Gold Cards and Sunflower Cards have been reduced by 15 % and 7 %, respectively. By analyzing customers' transaction records, potential small and micro corporate customers can be effectively identified. By utilizing remote banking and the cloud referral platform to implement cross-selling, considerable performance gains were achieved.

Obviously, the most classic application is in e-commerce. Tens of thousands of transactions are conducted in Taobao and the corresponding transaction time, commodity prices, and purchase quantities are recorded every day. More important, such information matches age, gender, address, and even hobbies and interests of buyers and sellers. Data Cube of Taobao is a big data application on the Taobao platform, through which, merchants can be ware of the macroscopic industrial status of the Taobao platform, market conditions of their brands, and consumers' behaviors, etc., and accordingly make production and inventory decisions. Meanwhile, more consumers can purchase their favorite commodities with more preferable prices.

The credit loan of Alibaba automatically analyzes and judges if to lend loans to enterprises through the acquired enterprise transaction data by virtue of big data technologies, while manual intervention does not occur in the entire process. It is disclosed that, so far, Alibaba has lent more than RMB 30 billion Yuan, with the rate of bad loans of only about 0.3 %, which greatly lower than those of other commercial banks.

6.3.2 Application of IoT Based Big Data

Internet of Things is not only an important source of big data, but also the main market of application of big data. In Internet of Things, every object in the real world may be both the producer and consumer of data and, because of the high variety of objects, the applications of Internet of Things also evolve endlessly.

Logistic enterprises may have profoundly experienced with the application of big data of Internet of Things. Trucks of UPS are installed with sensors, wireless adapters, and GPS, so the Headquarter can track truck positions and prevent engine failures. Meanwhile, this equipment also help UPS supervise and manage its employees, and optimize delivery routes. The optimal delivery routes specified by UPS for trucks are derived from their past driving experience. In 2011, UPS drivers have driven for nearly 48.28 million km less.

Smart city is a hot research area based on the application of Internet of Things data. The U.S. Miami-Dade County is a sample of smart city. The smart city project cooperation between Miami-Dade County in Florida and IBM closely connects 35 types of key county government departments and Miami City, and helps government leaders obtain better information support in decision making for managing water resources, reducing traffic jam, and improving public safety. IBM provides Dade County with smart instrument panel application by virtue of the in-depth analysis under cloud computing, so as to help the departments of county government with coordination-based and visualized management. The application of smart city brings about benefits in many aspects for Dade County. For example, Department of Park Management of Dade County saved one million USD in water bills due to timely identifying and fixing water pipes that were running and leaking this year.

6.3.3 Application of Online Social Network-Oriented Big Data

Online SNS is a social structure constituted by social individuals and connections among individuals based on an information network. Big data of online SNS mainly comes from instant messages, online social, micro blog, and shared space, etc. Since the big data of online SNS represents various user activities, the analysis of such data receives more attention. The analysis of big data of online SNS uses computational analytical method provided for understanding relations in the human society by

virtue of theories and methods, which involves mathematics, informatics, sociology, and management science, etc., from three dimensions including network structure, group interaction, and information spreading. The application of big data of online SNS includes network public opinion analysis, network intelligence collection and analysis, socialized marketing, government decision-making support, and online education, etc. Figure 6.1 illustrates the technical framework of the application of big data of online SNS. Classic applications of big data of online SNS are introduced in the following, which mainly mine and analyze content information and structural information to acquire values.

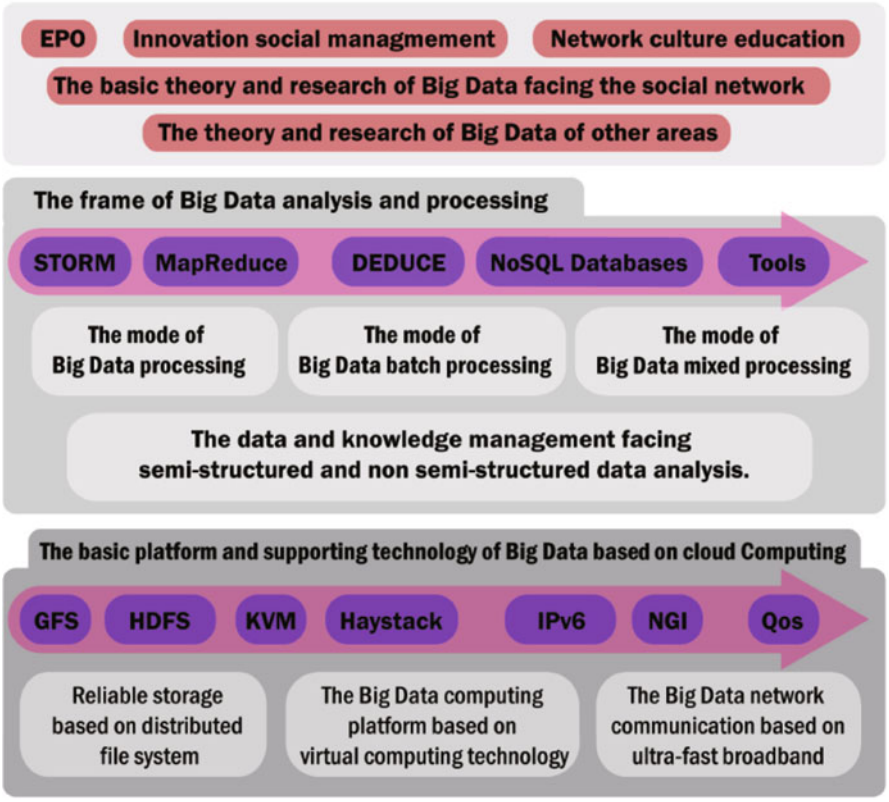

Fig. 6.1 Enabling technologies for online social network-oriented big data

- *Content-Based Applications*: Language and text are two most important forms of representation in SNS. Through the analysis of language and text, user preferences, emotions, interests, and demands, etc. may be revealed.
- *Structure-Based Applications*: On SNS with users as nodes, social relation, interest, and hobbies, etc. aggregate relations among users into a clustered

structure. Such structure with close relations among internal individuals but loose externally relations is also called a community. The community-based analysis is of vital importance to improve information propagation and for the research on interpersonal relation analysis.

The U.S. Santa Cruz Police Department experimented by applying data to conducting predictive analysis. By analyzing SNS, the police department can discover crime trends and crime modes, and even predict the crime rates in major regions [57].

In April 2013, Wolfram Alpha, a computing and search engine of the U.S., studied the law of social behaviors of users by analyzing social data of more than one million American users of Facebook. According to the analysis, it was found that most users of Facebook fall in love in their early 20s, get engaged when they are about 27 years old, get married when they are about 30 years old, and have slow changes in their marriage relationship between 30 and 60 years old. Such research results are highly consistent with the demographic census data of the U.S.

Global Pulse conducted a research that revealed some laws in social and economic activities using SNS data. This project utilized publicly available Twitter messages in English, Japanese, and Indonesian from July 2010 to October 2011, to analyze topics related to food, fuel, housing, and loan. The goal is to better understand public behavior and concerns. This project analyzed SNS big data from several aspects: predicting the occurrence of abnormal events by detecting the sharp growth or drop of the amount of topics; observing the weekly and monthly trends of dialogs on Twitter; developing models for the variation in the level of attention on specific topics over time; understanding the transformation trends of user behavior or interest by comparing ratios of different sub-topics; and predicting trends with external indicators involved in Twitter dialogues. As a classic example, the project discovered that the rice price follows the change of food price inflation from the official statistics of Indonesia, by analyzing topics related to rice price on Twitter (Fig. 6.2).

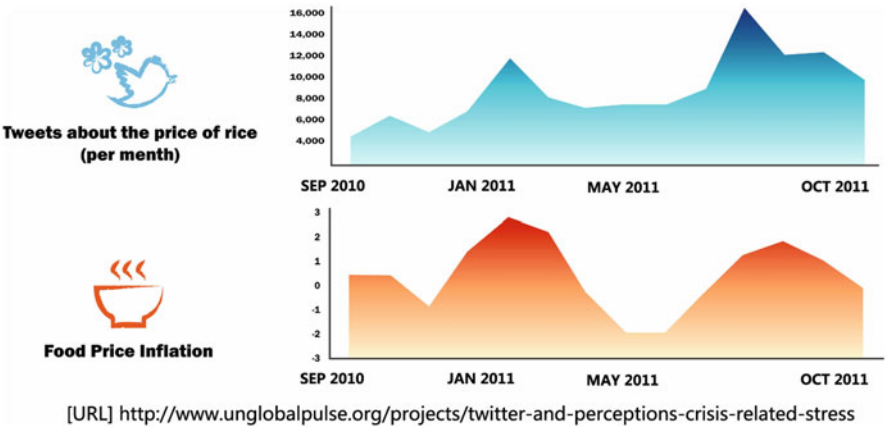

Fig. 6.2 The correlation between Tweets about the price of rice and food price inflation

Generally speaking, the application of big data of online SNS may help to better understand people's behavior and master the laws of social and economic activities from the following three aspects:

- *Early Warning*: to rapidly cope with crisis if any by detecting abnormalities in the usage of electronic devices and services.
- *Real-Time Monitoring*: to provide accurate information for the formulation of polices and plans by monitoring the current behavior, emotion, and preference of users.
- *Real-Time Feedback*: acquire groups' feedbacks against some social activities based on real-time monitoring.

The application of big data of online SNS involves three core technical problems:

- *Data Model*: Most traditional SNS data models are based on the static mode and specific analytical algorithms, and are not amenable for effective computation with data in the PB and higher scales. On the other hand, SNS analysis usually implements multi-dimensional complex relevant analysis on dynamic data. New theories and models need to be investigated to bridge this gap.
- *Data Storage and Management*: The existing Internet based storage management methods mainly support big data storage and rapid query. However, the existing approach does not effectively support the analytical computation of big data of online SNS, featuring high correlation, dynamic variability, and multi-dimensional evolution, etc. Therefore, new storage and management methods need to be developed.
- *Data Analysis*: The existing analytical methods on big data of SNS are mainly based on single-dimensional attribute, with insufficient accuracy. On the other hand, SNS analysis, such as topic evolution, group interaction, and public emotion drifting, etc., usually incorporates complex correlation analysis from the perspective of structure, group, and information. There is a need for the basic theory and methods to support complex correlation, multi-dimensional, large-scale, dynamic data.

6.3.4 Applications of Healthcare and Medical Big Data

Medical data is continuously and rapidly growing containing abundant and various information values. Big data has unlimited potential for effectively storing, processing, querying, and analyzing medical data. The application of medical big data will profoundly influence the human health.

For example, Aetna Life Insurance Company selected 102 patients from a pool of a 1,000 patients to complete an experiment in order to help predict the recovery of patients with metabolic syndrome. In an independent experiment, it scanned 600,000 laboratory test results and 180,000 claims through a series of detection test

results of metabolic syndrome of patients in three consecutive years. In addition, it summarized the final result into an extreme personalized treatment plan to assess the dangerous factors and main treatment plans of patients. This way, doctors may reduce morbidity by 50 % in the next 10 years by prescribing statins and helping patients to lose weight by five pounds, or suggesting patients to reduce the total triglyceride in their bodies if the sugar content in their bodies is over 20 %.

The Mount Sinai Medical Center in the U.S. utilizes technologies of Ayasdi, a big data company, to analyze all genetic sequences of Escherichia Coli, including over one million DNA variants, to know why bacterial strains resist antibiotics. Ayasdi's technology uses Topological data analysis, a brand-new mathematic research method, to understand data characteristics. HealthVault of Microsoft is an excellent application of medical big data launched in 2007. The goal is to manage individual health information in individual and family medical equipment. Presently, health information can be entered and uploaded with mobile smart devices and imported into individual medical records by a third-party agency. In addition, it can be integrated with a third-party application with the software development kit (SDK) and open interface.

6.3.5 Collective Intelligence

With the rapid development of wireless communication and sensor technologies, mobile phones and tablet computers have integrated more and more sensors, with increasingly stronger computing and sensing capacities. As a result, crowd sensing is coming to the center stage of mobile computing. In crowd sensing, a large number of general users utilize mobile devices as basic sensing units to conduct coordination with mobile networks for distribution of sensed tasks and collection and utilization of sensed data. The goal is to complete large-scale and complex social sensing tasks. In crowd sensing, participants who complete complex sensing tasks do not need to have professional skills. Crowd sensing modes represented by Crowdsourcing has been successfully applied to geotagged photograph, positioning and navigation, urban road traffic sensing, market forecast, opinion mining, and other labor-intensive applications.

Crowdsourcing, a new approach for problem solving, takes a large number of general users as the foundation and distributes tasks in a free and voluntary way. Crowdsourcing can be useful for labor-intensive applications, such as picture marking, language translation, and speech recognition. The main idea of Crowdsourcing is to distribute tasks to general users and to complete tasks that users could not individually complete or do not anticipate to complete. With no need for intentionally deploying sensing modules and employing professionals, Crowdsourcing can broaden the sensing scope of a sensing system to reach the city scale and even larger scales.

As a matter of fact, Crowdsourcing has been applied by many companies before the emergence of big data. For example, P & G, BMW, and Audi improve

their R & D and design capacities by virtue of Crowdsourcing. In the big data era, Spatial Crowdsourcing becomes a hot topic. The operation framework of Spatial Crowdsourcing is shown as follows. A user may request the service and resources related to a specified location. Then the mobile users who are willing to participate in the task will move to the specified location to acquire related data (such as video, audio, or pictures). Finally, the acquired data will be send to the service requester. With the rapid growth of usage of mobile devices and the increasingly complex functions provided by mobile devices, it can be forecasted that Spatial Crowdsourcing will be more prevailing than traditional Crowdsourcing, e.g., Amazon Turk and Crowdflower.

6.3.6 Smart Grid

Smart Grid is the next generation power grid constituted by traditional energy networks integrated with computation, communications and control for optimized generation, supply, and consumption of electric energy. Smart Grid related big data are generated from various sources, such as (a) power utilization habits of users, (b) phasor measurement data, which are measured by phasor measurement unit (PMU) deployed national-wide, (c) energy consumption data measured by the smart meters in the Advanced Metering Infrastructure (AMI), (d) energy market pricing and bidding data, (e) management, control and maintenance data for the devices and equipment in the power generation, transmission and distribution networks (such as Circuit Breaker Monitors and transformers). Smart Grid brings about the following challenges on exploiting big data.

- *Grid Planning*: By analyzing data in Smart Grid, the regions can be identified that have excessive high electrical load or power outage frequencies. Even the transmission lines with high failure possibility can be predicted. Such analytical results may contribute to grid upgrading, transformation, and maintenance, etc. For example, researchers from University of California, Los Angeles designed an "electric map" according to the big data theory and made a California map by integrating census information and real-time power utilization information provided by electric power companies. The map takes a block as a unit to demonstrate the power consumption of every block at the moment. It can even compare the power consumption of the block with the average income per capita and building types, so as to obtain more accurate power usage habits of all kinds of groups in the community. This map provides effective and visual load forecast for city and power grid planning. Preferential transformation on power grid facilities in blocks with high power outage frequencies and serious overloads may be conducted, as displayed in the map.
- *Interaction Between Power Generation and Power Consumption*: An ideal power grid shall balance power generation and power consumption. However, the traditional power grid is constructed based on one-directional approach of

transmission-transformation-distribution-consumption, which could not adjust the generation capacity according to the demand of power consumption, thus leading to electric energy redundancy and waste. To this end, smart electric meters are developed to enable the interaction between power consumption and power generation, and to improve power supply efficiency. TXU Energy has widely deployed smart electric meters with a big success. Power supply companies can read power utilization data every other 15 min other than every month in the past. Therefore, labor cost for meter reading is saved and, because power utilization data (a source of big data) are frequently and rapidly acquired and analyzed, power supply companies can adjust the electricity price according to peak and low periods of power consumption. TXU Energy utilized such price lever to stabilize the peak and low fluctuations of power consumption. As a matter of fact, the application of big data in the smart grid can help the realization of time-sharing dynamic pricing, which is a win-win situation for both energy suppliers and users.

- *Access of Intermittent Renewable Energy*: At present, many new energy resources, such as wind energy and solar energy, are also accessed to power grids. However, since the power generation capacities of such new energy resources are closely related to climate conditions that feature randomness and intermittency, it is challenging to access them to power grids. If the big data of power grids is effectively analyzed, such intermittent renewable new energy sources can be effectively managed: the electricity generated by the new energy resources can be allocated to regions with electricity shortage. Such energy resources can complement the traditional hydropower and thermal power generations.

References

1. Rita L Sallam, James Richardson, John Hagerty, and Bill Hostmann. Magic quadrant for business intelligence platforms. *Gartner Group, Stamford, CT*, 2011.
2. Beyond the pc. Special Report on Personal TEchnology, 2011.
3. Stephen A Goff, Matthew Vaughn, Sheldon McKay, Eric Lyons, Ann E Stapleton, Damian Gessler, Naim Matasci, Liya Wang, Matthew Hanlon, Andrew Lenards, et al. The iplant collaborative: cyberinfrastructure for plant biology. *Frontiers in plant science*, 2, 2011.
4. D Agrawal, P Bernstein, E Bertino, S Davidson, U Dayal, M Franklin, J Gehrke, L Haas, A Halevy, J Han, et al. Challenges and opportunities with big data. a community white paper developed by leading researchers across the united states, 2012.
5. George K Baah, Alexander Gray, and Mary Jean Harrold. On-line anomaly detection of deployed software: a statistical machine learning approach. In *Proceedings of the 3rd international workshop on Software quality assurance*, pages 70–77. ACM, 2006.
6. Michael Moeng and Rami Melhem. Applying statistical machine learning to multicore voltage & frequency scaling. In *Proceedings of the 7th ACM international conference on Computing frontiers*, pages 277–286. ACM, 2010.
7. Mohamed Medhat Gaber, Arkady Zaslavsky, and Shonali Krishnaswamy. Mining data streams: a review. *ACM Sigmod Record*, 34(2):18–26, 2005.

8. Vassilios S Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, 33(1):50–57, 2004.
9. Wil van der Aalst. Process mining: Overview and opportunities. *ACM Transactions on Management Information Systems (TMIS)*, 3(2):7, 2012.
10. Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha, and Horst D Simon. Pagerank, hits and a unified framework for link analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 353–354. ACM, 2002.
11. Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.
12. Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.
13. Yanpeng Li, Xiaohua Hu, Hongfei Lin, and Zhiahi Yang. A framework for semisupervised feature generation and its applications in biomedical literature mining. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(2):294–307, 2011.
14. David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
15. Helen Balinsky, Alexander Balinsky, and Steven J Simske. Automatic text summarization and small-world networks. In *Proceedings of the 11th ACM symposium on Document engineering*, pages 175–184. ACM, 2011.
16. Meenakshi Mishra, Jun Huan, Said Bleik, and Min Song. Biomedical text categorization with concept graph representations using a controlled vocabulary. In *Proceedings of the 11th International Workshop on Data Mining in Bioinformatics*, pages 26–32. ACM, 2012.
17. Jian Hu, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. Enhancing text clustering by leveraging wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 179–186. ACM, 2008.
18. Mark T Maybury. *New directions in question answering*. AAAI press Menlo Park, 2004.
19. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1–2):1–135, 2008.
20. Sankar K Pal, Varun Talwar, and Pabitra Mitra. Web mining in soft computing framework: Relevance, state of the art and future directions. *Neural Networks, IEEE Transactions on*, 13(5):1163–1177, 2002.
21. Soumen Chakrabarti. Data mining for hypertext: A tutorial survey. *ACM SIGKDD Explorations Newsletter*, 1(2):1–11, 2000.
22. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.
23. David Konopnicki and Oded Shmueli. W3qs: A query system for the world-wide web. In *VLDB*, volume 95, pages 54–65, 1995.
24. Soumen Chakrabarti, Martin Van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11):1623–1640, 1999.
25. Duo Ding, Florian Metze, Shourabh Rawat, Peter Franz Schulam, Susanne Burger, Ehsan Younessian, Lei Bao, Michael G Christel, and Alexander Hauptmann. Beyond audio and video retrieval: towards multimedia summarization. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 2. ACM, 2012.
26. Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys (CSUR)*, 44(4):25, 2012.
27. Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):1–19, 2006.

28. Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797–819, 2011.
29. You-Jin Park and Kun-Nyeong Chang. Individual and group behavior-based customer profile model for personalized product recommendation. *Expert Systems with Applications*, 36(2):1932–1939, 2009.
30. Ana Belén Barragáns-Martínez, Enrique Costa-Montenegro, Juan C Burguillo, Marta Rey-López, Fernando A Mikic-Fonte, and Ana Peleteiro. A hybrid content-based and item-based collaborative filtering approach to recommend tv programs enhanced with singular value decomposition. *Information Sciences*, 180(22):4290–4311, 2010.
31. Milind Naphade, John R Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *Multimedia, IEEE*, 13(3):86–91, 2006.
32. Zhigang Ma, Yi Yang, Yang Cai, Nicu Sebe, and Alexander G Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 469–478. ACM, 2012.
33. Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569, 2005.
34. Duncan J Watts. *Six degrees: The science of a connected age*. WW Norton & Company, 2004.
35. Charu C Aggarwal. *An introduction to social network data analytics*. Springer, 2011.
36. Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011.
37. Akira Ninagawa and Koji Eguchi. Link prediction using probabilistic group models of network structure. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1115–1116. ACM, 2010.
38. Daniel M Dunlavy, Tamara G Kolda, and Evrim Acar. Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(2):10, 2011.
39. Jure Leskovec, Kevin J Lang, and Michael Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web*, pages 631–640. ACM, 2010.
40. Nan Du, Bin Wu, Xin Pei, Bai Wang, and Liutong Xu. Community detection in large-scale social networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 16–25. ACM, 2007.
41. Sanchit Garg, Trinabh Gupta, Niklas Carlsson, and Anirban Mahanti. Evolution of an online social aggregation network: an empirical study. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 315–321. ACM, 2009.
42. Miltiadis Allamanis, Salvatore Scellato, and Cecilia Mascolo. Evolution of a location-based online social network: analysis and models. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 145–158. ACM, 2012.
43. Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 131–144. ACM, 2012.
44. Elena Zheleva, Hossam Sharara, and Lise Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1007–1016. ACM, 2009.
45. Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM, 2009.

46. Yanhua Li, Wei Chen, Yajun Wang, and Zhi-Li Zhang. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 657–666. ACM, 2013.
47. Theodoros Lappas, Kun Liu, and Evimaria Terzi. Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 467–476. ACM, 2009.
48. Tong Zhang, Alexandrin Popescul, and Byron Dom. Linear prediction models with graph regularization for web-page categorization. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 821–826. ACM, 2006.
49. Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2(1):718–729, 2009.
50. Wenyuan Dai, Yuqiang Chen, Gui-Rong Xue, Qiang Yang, and Yong Yu. Translated learning: Transfer learning across different feature spaces. In *Advances in Neural Information Processing Systems*, pages 353–360, 2008.
51. Cisco Visual Networking Index. Global mobile data traffic forecast update, 2012–2017 http://www.cisco.com/en.US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html(Sonerişim:5May\T1\is2013), 2013.
52. Youngho Rhee and Juyeon Lee. On modeling a model of mobile community: designing user interfaces to support group interaction. *interactions*, 16(6):46–51, 2009.
53. Jiawei Han, Jae-Gil Lee, Hector Gonzalez, and Xiaolei Li. Mining massive rfid, trajectory, and traffic data sets. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 2. ACM, 2008.
54. Eugene Wu, Yanlei Diao, and Shariq Rizvi. High-performance complex event processing over streams. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 407–418. ACM, 2006.
55. Manoj K Garg, Duk-Jin Kim, Deepak S Turaga, and Balakrishnan Prabhakaran. Multimodal analysis of body sensor network data streams for real-time healthcare. In *Proceedings of the International Conference on Multimedia information retrieval*, pages 469–478. ACM, 2010.
56. Yubin Park and Joydeep Ghosh. A probabilistic imputation framework for predictive analysis using variably aggregated, multi-source healthcare data. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 445–454. ACM, 2012.
57. Viktor Mayer-Schönberger and Kenneth Cukier. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt, 2013.