International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.6 (2015), pp.123-130
http://dx.doi.org/10.14257/ijmue.2015.10.6.12

Data Integration and Mining based on Web Big Data

Su-Zhi Zhang, Xu-Kai Qu and Jia-bin Sun

College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China

*zhangsuzhi@zzuli.edu.cn,282169146@qq.com,*258945405@qq.com

Abstract

As the revolution of Web 2.0 technology, more and more novel service industries, such as social network, web of things and mobile internet emerge. The data of Web explosive growth is called the "big data", which is hottest. Because of the great value of big data of Web, how to achieve the Web data and how to mine and utilize it, the two are paid attention by an increasing number of researchers. Under the big data circumstance, the Web data is characterized by huge scale, various kinds and high-speed bitstream. Therefore, we can investigate, further, the Web data mining, integration, interpretation and analysis. Simultaneously, Web data mining and integration still confront challenges consist of data scale, data variety, data timeliness and protection of privacy.

Keywords: *big data, Web data, integration and mining*

1. Introduction

With the rapid development of technologies about social network, Internet of things, mobile internet and Web 2.0, the explosive growth of Web data is gradually channeling from "massive data" to "big date". Because of the huge value of the big date itself, more and more institutions and researchers are beginning to pay attention to "big date" provided by the web [1]. Web date format is complicated, including structured, semi-structured and unstructured in variety. On one hand， a large number of users have a positive desire level of personal information，expecting to mining the value deeper, On the other hand, a large number of web data leave users wasting their time in choice, thus causing useless work increased and the satisfaction from users lowered [2]. Therefore, it is the most urgent problem to be solved on how to obtain, mine and use Web data today.

This thesis, based on the concept of Web data, will explore Web date processing over the past years, discuss the key technology of data integration and mining, analyze the status of Web data extraction and integration from Web data to Deep Web data, [3]list and introduce a multiple of hot field applications of data analysis, concerning natural language processing, social computing, recommendation system etc, and discuss the value and method of data interpretation. Finally, the challenge of big Web data mining and integration is also elaborated.

2. Web Data Mining and Integrating

The massive Web data mining and integrating can be regarded as a kind of typical big data application. Generally speaking, it can be divided into data integration and extraction as well as date analysis [4]. The whole specific processing process can be summarized as follows: extracting and integrating the widely-heterogeneous data source with the assistance of proper tools, inducing and arranging the result In line with a certain specification and analyzing the stored data utilizing appropriate data-analyzing techniques to extract the knowledge required and exhibit it to terminal users in an appropriate way.

ISSN: 1975-0080 IJMUE

2.1Data Mining and Integrating

2.1.1.Web Data: Web data's multi-agent interaction, cross-media correlation and strong real-time correlation make it a new problem confronted with big web data concerning Information extraction，data integration, data analysis and data interpretation [5]. The source of big data is extremely wide and data type is extremely diversified，accordingly, to process the big data, firstly, we should make data extraction from the data source, extract the entities and correlations and regulate the data stored with the help of integration and correlation in a unified structure form [6]. The data need screening in the integration and extraction to ensure credibility and authenticity of it.

Web Information extraction is prerequisite of big data processing, the method of which mainly including Information extraction based on Webpage structure, E Information extraction based on wrapper inducing, Information extraction based on pattern matching, Information extraction based on identity, etc.[7].

The characteristic of Information extraction based on Webpage structure is in line with structural location information of a web page [8]. There are many systems using this technology, such as commercial Lixto, non-commercial XWRAP, in addition, RoadRunner and W4F and SG-WRAM use the same technology. Moreover, Many scholars solve Web information extraction by using XML technology, developing a Web query prototype system based on XML.

Nicholas Kushmerick first put forward the Information extraction based on wrapper inducing in 1996, the method can automatically analyze the structural characteristics in a webpage and thus achieving information extraction, and the main idea of it is to generate extracting rules by taking the inductive learning method.[9, 10]. Due to Webpage structure uncertainty and volatility, webpage structural data extraction wrappers mostly adopt the implementation procedure of pre-assigned pattern.

Pattern matching is a highly adaptive method of information extraction system, mainly involving related techniques and approaches of natural language. There has several developed pattern matching systems, such as Simint, Cupid, LSD, iMAP and matching approaches based on copies, etc. In order to solve the problem of large query space of candidate matching generated in matching approaches at 1:1, Liu Shunjiang put forward the CBCSM matching method, which can enormously reduce the candidate matching query space, and improve the recall ratio and precision ratio at the same time effectively [3].

The core of information extraction based on identity is to construct identity, make use of the defined concept, Hierarchical classification, relation, function, axiom, cases and some essential external data to make information extraction in web pages, obtain a structured knowledge and save it. The identity crawler system, which gears search strategies to its users and domain identities, can effectively improve the accurate rate and recall rate of the page query.

2.1.2.Deep Web Data: According to the depth of data, the network space can be divided into Surface Web and Deep Web (or Hidden Web). Due to the sharp increase of Web data, it is rather difficult to obtain the information desired. To search more abundant, valuable and deep network-hidden information, a large number of researchers are making research into the Deep Web [11]. Deep web is a web page which fails to be detected by general search engines owing to technological factors or some information network contents which the search engine refuses to index for its own reason.

Now there exists more problems to be researched into in Deep Web data integration, mainly focusing on the discovery of Web database, the classification of web database, the extraction of query interface pattern, the integration of the query interface, the query transform, the extraction of the query result and the annotation of the result, etc. some of the problems have been widely studied [12]. Three common frameworks of Deep Web

data integration are seen: the module of query interface integration, query processing module and the module of query result processing.

At the same time, the uncertainty problem of Deep Web integration system also exists [6]. For a user, he may not have an exact idea what data to search and how to describe his query request. On the other hand, for a large scale dynamic integration Deep Web system, in a Web environment of high degree of autonomy, the more automatic it is, the more inaccurate it is. These uncertainties have become unavoidable in the system integration. Therefore, it is the focus of research on how to obtain high-quality data to improve the user experience and improve user satisfaction as a result.

Three uncertainty problems stand out in the Integration system, *i.e.* uncertain data, uncertain matching model and uncertain query processing. We can analyze these problems from three aspects. For those disorderly data in data layer, We can make adopt the visible page partition algorithm for cleaning and denoising, establish a secondary probability match mapping with the data source according to the interface attribute information and key word information in the mapping layer to achieve more relevant data and conduct a Top K reliability calculation to the query result in query layer to provide better data for the user.

Data come before patterns in the age of big data. Special attention should be paid to the relationship between data and patterns in data extraction and integration, because the pattern is in a constant evolution.

2.2 Data Analysis

Data analysis plays a decisive role in big data processing, and the value of big data is often generated in the process of data analysis, which is the key to realizing the value of big data. Data analysis is a process of collecting data in an intended way, analyzing data and making it information, the original data of which is extracted and integrated from the heterogeneous data source.

Web data analysis has been widely applied to the fields of natural language processing, data mining, recommending systems and the like.

2.2.1. Natural Language Processing: Natural language processing is one of the most important research directions in information technology, usually referring to the meaningful analysis and operation of natural human language by computer concerning lexical and syntactic analysis, semantic analysis, text classification and clustering, automatic summarization, text generation, sentiment analysis etc.. In recent years, various aspects of natural language processing have observed a more in-depth study.

In sentiment analysis, there is a new emotion recognition model based on linguistic structure of network reviews or fixed sentiment terms model, which constructs the recognition algorithm by utilizing three specific combinations patterns based on fixed emotion word elements, and constantly update feature word element set through the relevant user's feedback based on the incremental TF-IDF model. Compared with the traditional emotion recognition method, this method can significantly improve the efficiency and accuracy of sentiment classification.

In text classification, specific to the technical difficulties of large size of short text classification samples, sparse keywords feature and difficulty in processing technology, the KNN short text classification algorithm based on semantics adopts segmentation strategy based on word to extract feature words in short text, conducts a concept mapping of keywords combined with the CNKI to improve the semantic expression of short texts, and uses the LSA dimensionality reduction according to the characteristics of short texts, improves the KNN classification algorithm, thus effectively improving the classification performance of short texts. Compared with the traditional classifier, an improved KSVM classifier has greatly improved the classification accuracy of short texts, which can adopt different classifiers based on the misclassification problem by a support vector machine to

the treatment of samples near the hyper plane and the calculation of the sample to be identified and the distance of the optimal classification hyper plane in the classification phase. It is also applicable to short text classification.

Against the backdrop of big data era background, a series of new challenges is put forward concerning natural language processing: general uncertainty, covering lexical, syntactic, semantic, pragmatic and voice at all levels; the unpredictability of an unknown language phenomenon, including an endless emergence of new words, new terminology, new semantics and grammar; ever-confronted data insufficiency, that is, limited language collection always covers ever-growing language phenomena; the complexity of language knowledge expression, i.e. the fuzziness of semantic knowledge and perplexing relevance cannot be described by a simple mathematical model, the cultural connotation is difficult to be effectively described with conventional methods, and the semantic computing needs a nonlinear calculation of huge parameters.

2.2.2. Social Computing: With the arrival of the big data era, social computing, as a data-intensive science, has had a great effect on the scope, depth and scale of data collection and analysis. Social Computing mainly has its development trend in two aspects: one is a social science geared to the needs of the society like computational social science and social network analysis; the other is for technical applications, such as social networks, entertainment applications etc.

Among them, the explosive growth of social network users and user generated content information, have brought about huge challenges to the online social network analysis and community discovery. More and more scholars are beginning to pay attention to the efficiency and highly-accurate processing of the social network analysis [13]. Xiong Zhengli and some other people proposed a discovery algorithm based on the users of online social network community according to the unique characteristics of online social network and aiming at the hard-detected problems of potential communities for online social networks [14]. Fang Ping and some other people proposed a new discovery algorithm of community structure based on the number of common friends and neighbor nodes information. These methods, to a certain extent, have helped to improve the accuracy.

However, against the backdrop of big data, social computing is still in its primary stage, its development is not mature, and also exist some problems and challenges. The following aspects are mainly included:

The algorithm selection. The key issue is to select the appropriate algorithm and parallel strategies for data processing.

The credibility of software and service. The cloud environment highlights the consideration of privacy security.

Too much uncertainty.

2.2.3. Recommendation System: Recommendation system is one of the important information filtering mechanism, which is a very promising way to solve the problem of information overload. It can help users find and even find the information resources that they may need and be interested in from among a large amount of information in the network by mining the link law between the users and information resources, studying the interest and preferences of the users and carrying out a personalized computing to find out the interests of the users through the system, thus guiding the users to find their information needs [15, 16]. Based on all kinds of intelligent algorithms, the recommendation system extracts the indicators and patterns by mining mass data and generates personalized recommendation results.

Four main methods of data analysis for the recommendation system are observed: recommendation based contents, recommendation based on association rules, recommendation based on knowledge and recommendation based on collaborative

filtering, of which collaborative filtering recommendation is the most widely used at present, and recommendation algorithm is realized through the score prediction of unrated items [15]. The Recommendation system based on collaborative filtering was first put forward by Goldberg and some others in 1992, finding its application in Typestry system and the target users need to point out other users relatively similar to them in behavior.

Against the deficits of user interest and data sparsity for the filter algorithm of the existing system, a collaborative filtering recommendation algorithm based on user interest thus emerges as required. The algorithm introduces weight of user interest, establishes an evaluation matrix for user items, finds out interested users similar to specified users from among the user groups, carries out a similarity calculation, forms the prediction of preferences for the specified user by information system, thus getting the recommended result [17]. The experimental result shows that the algorithm can filter based on some complex and difficult-to-express concepts (the information quality, grade) and improve the quality of recommendation by exploiting the user interest.

At present, most of the recommender systems use collaborative filtering recommendation technology for Web log mining and generates a recommendation result [11]. The algorithm focuses on the accuracy of recommendation yet ignores the diverse recommendation. The sorting-based diversity recommendation algorithm, based on the improved collaborative filtering algorithm on AB neighbor, uses (LCM and I-tree) classification algorithm to sort the recommendation results, thus improving the diversity of recommendation obviously while ensuring the accuracy of recommendation at the same time.

The rapid development of Web2.0 in recent years makes the problems of information-overloaded social networking site and e-commerce more and more serious, and the use of traditional recommendation system has failed to meet the needs of users. Mikofi and some others mainly discussed in the "Open source recommendation system analysis of big data" the recommendation system for big data sets, analyzed the existing open source recommendation system Mahout, Duine and EasyRec, and then described how open source recommendation system could meet the changing challenges of big data age through the evaluation index (and definition 4V of big data).

Mahout is an open source project under Apache software Foundation, which is a distributed framework of machine learning and data mining, aimed to help developers create intelligent applications more easily and quickly [5]. The greatest advantage of Mahout is being able to be efficiently extended to cloud computing framework by using the Apache Hadoop library. It converted some PC-based algorithm to MapReduce mode, improving the processing performance of the algorithm. The current commonly-used algorithms like recommendation algorithm, clustering algorithm and classification algorithm all have very good realization.

2.2.4. Data Interpretation: The interpretation of data is to show the result of data analysis in an easily-understood way. Generally, the traditional practice of the interpretation of data is to output the result in the form of text, or display directly on the computer terminal. When the data is relatively small, the traditional way can narrowly meet the needs of users, but in the age of big data, not only the input of data analysis is massive, also its output is unavoidably massive, and at the same time, a great complexity exists among the output results. Hence, using the adoption of the traditional interpretation method is poor in feasibility and effect.

The ability of data interpreting can be enhanced by us from two aspects: visual technology and human-computer interaction. Data Visual technology is the theory, technique and technology resorting to the use of computer graphics and image processing technology to convert data into graphics or images and then display them on the screen and conduct an interactive processing [9]. Interactive technology, as it is called, is allow users to understand and participate in the specific process of analysis to a certain extent.

Concerning the visualization of data analysis, the domestic web of Renren has launched the service of friends' archives (still in the development stage), provided online analysis service based on the information of the user's friends, presented the male-to-female ratio of the friends, university distribution, region distribution map, statistical surnames, the attention to the current user through statistical analysis and graphical presentation, helping people to know more about their friends available and thus enhancing the interaction between friends.[5]

3. Challenges Big Date Integrating and Mining Encounter

3.1 Data Scale Exponentially Growing

Web data make people encounter unprecedented large scale of data in extraction, integration, data analysis and interpretation of data. Here not only refers to the huge amount of data, data from GB, TB, the magnitude of PB development to EB, or even ZB (zettabytes, equal to 270 bytes), also represents the high complexity of Web large data, also have to face more complex data objects, typical characteristics for a variety of modes and types of diversity the relationship between the complex, uneven quality, correlation. [9]At the same time, bottleneck problem appears in coordination of data and the corresponding storage device, because the traditional database the pursuit of a high degree of data consistency and fault tolerance, extended availability issues lack of Data types are becoming more diverse, is challenging the traditional data analysis platform. From a database perspective, elastic and effective mining algorithm is the key to the realization of data mining, small data and the current algorithm and systematic, data effect of stored video, audio and other non-structured and semi-structured.

At present, data storage capacity has been the continuing growth of the size of the data far behind, can not keep up the pace, so the design reasonable and effective hierarchical distributed storage architecture becomes the key web data management.

3.2 The Diversity of Data Types

The age of big data, fusion data types from structured data to structured and half-structured and non-structured types. Suitable for memory resident datasets, data in large databases at the same time into memory is very difficult, as the data size increases with each passing day, a high efficiency algorithm has gradually become the bottleneck of the data analysis process. In order to completely change the passive situation, we need to restructure the existing framework, organization system, resource allocation and power structure.

3.3 The Timeliness of Large Web Data Processing

Data size increase makes the data processing time becomes longer and longer, and large data condition on the timeliness of information processing requirements higher and higher. [2]Now every day in the Web data to produce the massive, sometimes the need for real time processing of these data, to the mining of the information, such as public opinion monitoring, real time data processing is quite a challenging job, the data stream itself has continued to reach, speed and large scale, so the usual practice does not for permanent storage of all data, coupled with the data environment is changing continuously, the system to accurately grasp the panorama of the whole data quite difficult. Study on the theory and techniques for data flow is still of great research value in large data. At the same time many practical systems will also be data stream technology were developed and widely applied, representative open source system such as Twitter Storm, Yahoo S4 and Linked in Kafka etc.

3.4 Privacy Protection

Because of the personal information and the activities of people's increasing trajectory in the network, the Web data value has not only become increasingly prominent, but also breed a number of lawless elements, steal user information to make against the user and national collective behavior. This makes the big data safety very important [1]. Facing with safety and privacy protection network data, there are a lot of problems need to be addressed in, including: data computing ethics, data cryptography, distributed programming framework of secure computing, remote data calculation of reliability, data storage and log management security, based on the mining and analysis, privacy and commercial interests of protection data the mandatory access control and secure communication, multi granularity access control and data sources and data channels credibility.

4. Conclusion

The sustainable development of internet generates the big web data. The thesis raises the basic concept of big data, makes a detailed analysis of key technique for the management of web data, elaborates the challenges confronted with web data management from big data and research findings available for big web data management and emphasizes some challenges facing big data. Although big data is not a new concept, the solution to the management of big Web data remains a long way to go.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.61201447.

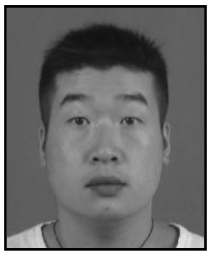
References

[1] E. Isikli, A. Ustundag and E. Cevikcan, "The Effects of Environmental Risk Factors on City Life Cycle: A Link Analysis [J]", Human and Ecological Risk Assessment, vol. 21, no. 5, (**2015**), pp. 1379-94.
[2] Z. Huang Shu, S. Li Tao and D. Luo, "Research on Algorithm for Mining Frequent Itemsets Based on Cloud Computing", In: Yarlagadda P, Kim YH, editors. Measurement Technology and Its Application, Pts 1 and 22013. p. 1303-7.
[3] A. Mozo-Villarias, J. Cedano and E. Querol, "A Model of Protein Association Based on Their Hydrophobic and Electric Interactions [J]", Plos One, vol. 9, no. 10, (**2014**).
[4] L. Vu and G. Alaghband, "Novel parallel method for association rule mining on multi-core shared memory systems [J]", Parallel Computing, vol. 40, no. 10, (**2014**), pp. 768-85.
[5] X. Zheng and S. Wang, "Study on the Method of Road Transport Management Information Data Mining Based on Pruning Eclat Algorithm and MapReduce [J]", 9th International Conference on Traffic and Transportation Studies (Ictts 2014), vol. 138, (**2014**), pp. 757-66.
[6] J. S. Yoo and D. Boulware, "A framework of Spatial Co-location Mining on MapReduce", Hu X, Lin TY, Raghavan V, Wah B, BaezaYates R, Fox G, et al., editors2013.
[7] Y. Wang, Z. Zhang and F. Wang, "A Parallel Algorithm of Association Rules Based on Cloud Computing [J]", 2013 8th International Icst Conference on Communications and Networking in China (Chinacom), IEEE, (**2013**), pp. 415-9.
[8] S. Sundaramoorthy and SP. Shantharajah, "AN IMPROVED ANT COLONY ALGORITHM FOR EFFECTIVE MINING OF FREQUENT ITEMS [J]", Journal of Web Engineering, vol. 3, nos. 3-4, (**2014**), pp. 263-76.
[9] S. Senhadji, S. Khiat and H. Belbachi, "Association Rule Mining and Load Balancing Strategy in Grid Systems [J]", International Arab Journal of Information Technology, vol. 11, no. 4, (**2014**), pp. 338-44.
[10] RB. Roberson III, TR. Elliott, JE. Chang and JN. Hill, "Exploratory Factor Analysis in Rehabilitation Psychology: A Content Analysis [J]", Rehabilitation Psychology, vol. 59, no. 4, (**2014**), pp. 429-38.
[11] 황정희, 신예호 and RK. Ho, "An Active Candidate Set Management Model on Association Rule Discovery using Database Trigger and Incremental Update Technique [J]", Journal of KIISE : Databases, vol. 29, no. 1, (**2002**), pp. 1-14.
[12] RW. Parks and J. Cardoso, "Parallel distributed processing and executive functioning: Tower of Hanoi neural networkmodel in healthy controls and left frontal lobe patients [J]", The International journal of neuroscience, vol. 89, nos. 3-4, (**1997**), pp. 217-40.

[13] S. Doddi, A. Marathe, SS. Ravi and DC. Torney, "Discovery of association rules in medical data [J]", Medical informatics and the Internet in medicine, vol. 26, no. 1, (**2001**), pp. 25-33.
[14] D. Apiletti, E. Baralis, T. Cerquitelli, S. Chiusano and L. Grimaudo, "SEARUM: a cloud-based SErvice for Association RUle Mining [J]", 2013 12th Ieee International Conference on Trust, Security and Privacy in Computing and Communications (Trustcom 2013), IEEE, (**2013**), pp.1283-90.
[15] N. Dang, V. Bay and L. Bac, "Efficient strategies for parallel mining class association rules [J]", Expert Systems with Applications, vol. 41, no. 10, (**2014**), pp. 4716-29.
[16] H-Y. Chang, S-C. Huang and Lai C-C, "A personalized IPTV channel-recommendation mechanism based on the MapReduce framework [J]", Journal of Supercomputing, vol. 69, no. 1, (**2014**), pp. 225-47.
[17] JT. Krogel, J. Kim and FA. Reboredo, "Energy density matrix formalism for interacting quantum systems: Quantum Monte Carlo study [J]", Physical Review B, (**2014**), vol. 90, no. 3.

Author

Suzhi Zhang, he received the Ph.D. degree in computer software and theory from Huazhong University of Science and Technology, Wuhan, China, in 2003. He is a senior member of China Computer Society, middle-aged backbone teacher and the key discipline of the first leader in computer application technology of Henan province. He published more than 60 papers in international academic journals and foreign academic conferences, 15 of these indexed by EI and ISTP. Since 2007, he has been with the faculty of the College of Computer and Communication Engineering, Zhengzhou University of Light Industry, where he is currently a Professor. His major research interests include Web data integration, distributed data base system, Social computing, Big data Management, and Mobile data privacy protection.



Xukai Qu, he received the B.S. degree in information engineering from Zhengzhou University of Light Industry, Zhengzhou, China, in 2011. He is currently pursuing the master's degree in computer technology at the School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, China. His current research interests include Data mining and integration, Web data mining, and Analysis of e-commerce data.

