2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)

Research on Technology, Algorithm and Application of Web Mining

Yeqing Li
Inner Mongolia University of Finance and Economics
Hohhot, China
lyqing4567@163.com

***Abstract*— With the rapid development of Internet, we have entered an era of information explosion, there is a lot of redundant information in the Network. How to extract a useful part of this information from the massive information resources, analyzing the vast amount of information and finally get the potential knowledge we want to extract. Web mining technology came into being, and saved out the human from the information ocean. This paper will analyze the realization of Web content mining and Web structure mining, their basic algorithm principles and their application areas.**

Keywords- Web mining technology; Web structure mining; Web content mining; Multi-media mining

I. INTRODUCTION

What is Web mining? It is the process that discover and extract the useful mode and knowledge that people are interested from the massive Web documents and activities through data mining technology [1]. Compared to the well-known Data mining, Web mining can be extended to a deeper and broader areas, the differences between them are also very obvious: the object of data mining is the data stored in database, that is to say, the structured data; Web Mining aims at the contents or structure of Web document, which has a feature of wide-distributed, dynamic and heterogeneous, and contains unstructured or semi-structured data.

Based on the diversity of information on the Web, Web mining is divided into the following category as shown in figure 1: Web structure mining, Web content mining and Web usage mining [2]. These three mining methods are different in the aspect of dealing with the main data, processing methods and application areas.

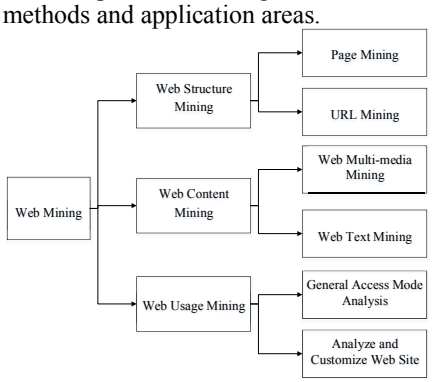

Figure 1. Web mining classification

1) Web Structure Mining mainly deals with Web structure data, it can be divided into page structure mining and URL mining (hyperlink mining) [3].

2) Web Content Mining mainly deals with unstructured data and semi-structured data, can be refined into Web text mining and Web multimedia mining based on the content, in which multimedia mining is a popular research topic at present [4].

3) Web Usage Mining can be divided into general access mode analysis and customizing Web site, it analyzes Web site logs to find some valuable knowledge.

This paper will analyze the realization of Web content mining and Web structure mining, their basic algorithm principles and their application areas.

II. WEB STRUCTURAL MINING

A. *Introduction of Structural Mining*

Massive Web sites constitute the entire Internet network, and each page in these Web sites more or less includes some hyperlinks, which contains a lot of potential information. The purpose of Web structure mining is to dig out the hidden knowledge, so that it can be fully applied. For example, if a paper is cited for a lot of times, it is proved that this paper is very authoritative in its field of study. Similarly, if there are a lot of Web pages pointing to the same page X, we think that page X has higher authority. In the field of search engine, it is very important to place the most authoritative page at the first of the search results, because when using the search engine users want to acquire authoritative and publicly recognized results, rather than some of the incorrect result pages or even insignificant ad page. Structural mining is based on this judging method to get a lot of information and help people with navigation and recommendations of the authoritative Web pages. This article will briefly introduce two well-known structure mining algorithm, PageRank and HITS.

B. *PageRank Algorithm*

The PageRank algorithm draws on the traditional citation analysis: when page A has a link to page B, we think that B gets the score that A contributes to it. This score depends on the importance of A, that is to say, the more important page A is, the higher the contribution score of page B gets. Based on this premise, Google uses this algorithm to give each page a number of values (PageRank) as a reference to the page quality. PageRank values from 0 to 10, the higher the value, the higher the quality and popularity of the page, the more forward the corresponding search results. The

978-1-5386-3221-5/17 $31.00 © 2017 IEEE
DOI 10.1109/CSE-EUC.2017.152

algorithm is not related to the user's query conditions, that is to say, the algorithm is irrelevant to the theme, it full uses PageRank value as a site evaluation criteria. This is the basic idea of the PageRank algorithm [5].

C. HITS Algorithm

HITS involves two important concepts: the content authority (Authority) -- the number of a certain Web page being referenced by other web pages; the link authority (Hub) -- the number of a certain page pointing to another page [6].

In the HITS algorithm, if a page's Authority is high, then the page is pointed by many other pages, which indicates that its content is of high quality; and if a page's Hub is high, then this page points to many other high-quality pages. When the search results are sorted at the end of the search, the pages are sorted from high to low according to the score of the Authority, and several pages with the highest weight are taken out as the search result of the response users' query. This is the basic idea of the HITS algorithm. The relationship between the Authority and the Hub makes the HITS algorithm able to dig into more authoritative pages.

D. The advantages and disadvantages of the PageRank algorithm and the HITS algorithm

It is undoubtedly that the PageRank and HITS are two classic algorithms based on hyperlink structure analysis for Web structure mining. But there are huge differences in the concrete realization process between these two algorithms, and thus have their own advantages and disadvantages.

The advantage of the PageRank algorithm is the use of off-line computing, which has a relatively short response time; and this algorithm excludes a large part of the human factors, only calculate its authority and importance through Web links between the sites. However, the PageRank algorithm uses the method of allocating the weights of the links evenly in the calculation, and may assign same authority to a highly related web page and a lowly related one. This ignores the actual situation of the link, and has its own irrationality [7].

A lot of practice has proved that HITS algorithm has a good precision ratio and recall ratio for many queries. But the HITS algorithm needs to complete the real-time calculation online, and the response speed is relatively slow; there are lots of pages whose links are irrelevant to the query topic but point to each other, HITS algorithm is likely to give such page a high ranking, resulting in Topic-Drift phenomenon.

E. Hilltop Algorithm

Rishna Baharat excogitated the HillTop algorithm at around 2000 and later authorized Google and helped it to complete a significant technical update in search ranking. HillTop adopts the basic guiding principle of PageRank, that is, determine the sorting weight of search results through the number and quality of link in the page, the difference is that HillTop use term input by user in the query box to determine the authority of a page [8].

Figure 2 is the flow chart of Hilltop algorithm, including two main processes which are expert page search and target page sorting.

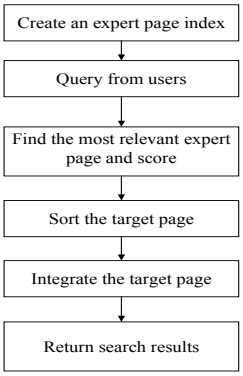

Figure 2. Hilltop algorithm flow chart

HillTop has the advantage of judging high-quality web pages with scientific and objective methods, which improve the search quality. But this algorithm also has many shortcomings, such as operational efficiency and scalability are relatively low, and cannot avoid some sites' deliberately cheating in order to improve the ranking. Now the HillTop algorithm and the PageRank algorithm have been combined to serve Google for better results.

III. WEB CONTENT MINING

A. Web Content Mining Introduction

Web Content Mining is a process of discovering valuable knowledge and obtaining useful potential information from massive Web data information. It can be divided into two types in terms of content, including text documents (text and hypertext data) and multimedia documents (video, audio, images, graphics and other multimedia data). Based on the above classification, content mining is also divided into two directions which are text mining and multimedia mining, text mining has always been a hot research topic, but in recent years, the study of multimedia mining has gradually been payed more attention from scholars.

B. Text Mining

Text Mining is the process of discovering and extracting implied knowledge from Web documents and eventually organizing the information that can be used directly by the user. The processed data are semi-structured and unstructured data, in which mainly include free text, HTML marked hypertext. The process of text knowledge discovery can be summarized as shown in figure 3: text preprocessing, text mining, pattern assessment and representation.

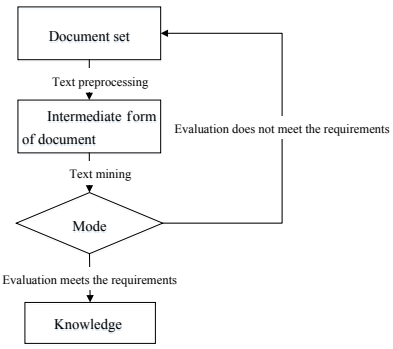

Figure 3. The process of text knowledge discovery

In terms of the function, the main techniques of text mining for the large number of Web documents include text summarization, text classification, text clustering, text association rule analysis and so on. The classification and clustering of text are the most important and basic in Web text mining. In terms of methods, text mining is divided into Information Retrieve method and Database method.

1. Information retrieve method and database method

The information retrieve method mainly uses the information query technology to evaluate and improve the quality of the search result information, and can also deal with unstructured data and HTML structure of semi-structured data, mainly used in text classification, clustering and pattern discovery; database methods and data warehouse method use data extraction and conversion method to convert or map the unstructured Web data information into structured data, and then mine the information with data mining technology.

2. Text summary

Text summary refers to presenting the core information from text in the form of brief description, so that users no longer need to click on the full text. The document summary can be used to the query result section of the search engine.

3. Text categorization

Text categorization is to identify the discriminant function through training on classification rules based on existing data, use this function to identify and classify a variety of unknown Web data with different attributes, so that it is more convenient and effective for users to search and read Web documents.

The current word segmentation method can be concluded into the mechanical segmentation and understanding segmentation. Currently, the most mature word segmentation should be Chinese lexical analysis system ICTCLAS, which is developed by the Chinese Academy of Sciences. There are many algorithms for text classification, the most important ones are Naive Bayesian classification algorithm, K-nearest neighbor algorithm, decision tree algorithm, neural network algorithm and support vector machine algorithm. This technology can automatically sort a large number of Web documents thus save a large amount of time. Text classification can be used for information retrieval, text filtering, document automatic classification, digital library and other fields, one of the important applications is to identify and filter spam massages and emails. From the current results, this technology has brought great convenience for our work and life.

4. Text clustering

Text clustering is an unsupervised induction machine learning problem. The existing text clustering algorithm can be roughly classified into two types: Hierarchical Clusters represented by G-HAC and Plane division algorithms represented by K-Means.

In computer programming, the steps of the K-means algorithm are as follows:

(1) Take K elements from the original data as each center of the K clusters.

(2) Calculate the minimum distance between other elements and the cluster center, assign these elements to the nearest cluster.

(3) According to the clustering results, recalculate the respective centers of the K clusters. The method is to calculate the arithmetic mean of the respective dimensions of all the elements in the cluster.

(4) Re-cluster all the elements according to the new cluster center.

(5) Repeat the previous step until the clustering result no longer changes, output the result.

The significance of text clustering is that it can divide a large number of text into several categories in accordance with its attribute or content, this greatly facilitates the user's search, and saves user's browsing time. As the text clustering does not require training process, so its flexibility and convenience has made it a major way to deal with Web documents. Text clustering can be applied to provide summary of large-scale documents' contents, identify similarity between hidden documents, ease the browsing of related information, etc.

5. Correlation Analysis

In addition to text classification and text clustering, there is an important method in text mining - Correlation Analysis. The correlation analysis is such an implication relationship:

$X \rightarrow Y$, of which $X \subset Y$, $Y \subset I$ and $X \cap Y = \emptyset$, X or Y is a collection of items, also called item-sets, X is the antecedent, Y is the consequent.

The support of $X \rightarrow Y$ refers to the percentage of $X \cup Y$ transactions included in collection T, which is the estimated of the probability Pr $(X \cup Y)$. If n is the number of transactions in T, the support of rule $X \rightarrow Y$ is:

$$\text{Support} = \frac{(X \cup Y).count}{n}$$

In which $(X \cup Y).count$ is the count of $X \cup Y$ in collection T, which is the transaction number of $X \cup Y$ in collection T.

The Confidence of $X \rightarrow Y$ refers to the percentage of number of the transactions containing both X and Y accounts for number of all transactions that contain only X, that is, the estimate of the conditional probability Pr (Y|X). The confidence degree of rule $X \rightarrow Y$ is:

$$\text{Confidence} = \frac{(X \cup Y).count}{X.count}$$

The task of associating rule is to find out all the rules that meet the pre-specified frequency and precision criteria. The most classic and basic method is the Apriori algorithm, and the hash-based approach, the partition-based method and the FP- Growth algorithm are all developed from the Apropri algorithm.

Apriori algorithm is a frequent item set algorithm that mining correlation rules. The core idea is to extract frequent item sets by two stages: candidate set generation and plot downward detection. The process consists of connection (class matrix operation) and pruning (remove those unnecessary intermediate results) [9].

The process of the Apriori algorithm is, when the number of items in the set is greater than 0:

(1) Construct a list of candidate items consisting of k items (k starts from 1).

(2) Calculate the support of the candidate set, delete the infrequent item set.

(3) Construct a list of candidate items consisting of k + 1 items.

(4) End of the cycle when the current frequent item set k only contains one item.

IV. MULTI-MEDIA MINING

Multi-media mining refers to extracting valuable knowledge from the mass multimedia data in the Web. Through a comprehensive analysis of audio-visual features and semantics, find implicit, effective, valuable, understandable model, and get the trend and relevance of the business, thus provide users with problem-solving level of decision support capacity [10].

Multimedia mining mainly involves two research areas which are data mining and multimedia information processing. From the perspective of processed data, it generally includes graphics and image data mining, video data mining and voice data mining. The process of multimedia mining can be roughly divided into three steps:

(1) Extract the required metadata from a large number of multimedia data and organize it into a meta-database;

(2) Use appropriate algorithmic techniques for mining multimedia content;

(3) Present and explain the mining results clearly.

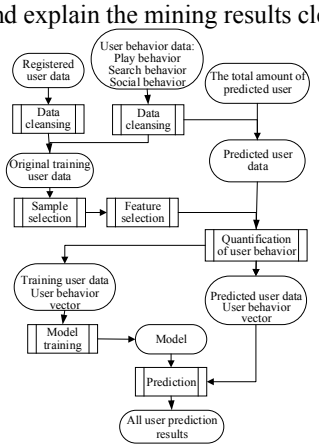

Figure 4. Multimedia data mining technology

As shown in figure 4, taking the video content as example, a series of data such as user registration information, playback record, search behavior, social behavior are analyzed, and the user's attributes and behavior habits are judged by cleaning, analysis and forecasting to complete the user's portrait, thus to achieve the purpose of precise advertising and recommending website content and other business purposes.

Multimedia mining involves contents of many areas, though difficult to research, it is a very promising area. Nowadays, multimedia content occupies a large proportion of Web contents, and has great mining value. There must be unlimited prospects if the multimedia mining technology can be applied in commercial field.

V. CONCLUSION AND FUTURE WORK

With the emergence and development of Web mining, this technology is not only used in the field of search engines, but also involves in e-commerce, online shopping, e-learning, e-government and other aspects of social life.

In addition, through Web mining technology, people have a new understanding of artificial intelligence, and has also made new breakthroughs network security. As a kind of technology to extract and discover knowledge from massive data, Web mining technology has become the basis of a large number of emerging Internet technologies, and has created imponderable value. Web mining technology is also experiencing constantly progress and update. With joint unremitting efforts from exports, the future of Web mining technology will be more developed, and eligible to help people to solve more problems.

REFERENCES

[1] Kosala R, Blockeel H. Web mining research: a survey[J]. Acm Sigkdd Explorations Newsletter, 2015, 2(1): 1-15.

[2] Cooley B R W. Web usage mining: Discovery and application of interest in patterns from web data[J]. Campus-Wide Information Systems, 2010, 24(5):308-330.

[3] Zewen Li. Web-based Data Mining Technology[J]. Modern computer, 2011,3(15):51-58

[4] Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine [J]. Computer networks, 2012, 56(18): 3825-3833.

[5] Rong Zhang. Research on Technology of Web Mining[J]. Computer Engineering: 2006.08 vol.32(15)

[6] R.Lempel and S.Moran, SALSA: stochastic approach for link-structure analysis and the TKC e ect[J]. ACM Trans, Information Systems: 2001,19:131-160

[7] Monika R.Henzinger and Krishna Bharat. Improved algorithms for topic distillation in a hyperlinked environment. Proeeedings of the 21’st Internationa1 ACM SIGIR Conference on Research and Development in IR, 1998.08

[8] Bishui Zhou, Yanhong Zhang. HillTop Algorithm analysis[J]. Computer age: 2005 vol.4

[9] Zhao Y. Association Rule Mining with R[J]. 2015

[10] Osmar R, Zaiane, Jiawei Han, Ze-Nian Li, Jean Hou. Mining Multimedia Data. CASCON' 98: Meeting of Minds, Toronto, Canada, November,1998.83-96.