The Research of Web Mining

Lizhen Liu, Junjie Chen, Hantao Song
Department of Computer
Beijing Institute of Technology
Beijing, 100081
China

***Abstract*—With the prompt increasing of information in the WWW, the Web Mining has gradually become more and more important in Data Mining. People always hope to gain some efficient knowledge patterns through searching, integrating, mining and analyzing on web. These useful knowledge patterns can help us to build the efficient web site that can serve people better. The researches of Text Mining and Usage mining on web are introduced in the text. In the end of the paper we gave an applicable example of Usage Mining.**

I. INTRUDUCTION

To discover and analyze useful information in the WWW by using data mining techniques has gradually become a important direction of knowledge discovery. The WWW serves as a huge, wide, distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other services[1]. The Web also contains rich and dynamic collection of hyperlink information, and usage information, providing rich sources for data mining. Web mining includes mining Web linkage structure, Web contents, and Web access patterns[2]. This involves mining Web link structures to identify authoritative Web pages, the automatic classification of web documents, and Web log mining.

But complex contents of Web page are different from traditional text documents. They haven't uniform structure. In addition, Web is a quickly change information source, not only in Web contents but also in page contents. For example, news, stock market, company advertisement, and network service center all modify their information on Web during some period. In addition, Web page linkages and accessing paths are always changed. Facing constantly increased user quantity, different interests and various motives of users, how can we obtain the information which user are interested in? How can we obtain high quality Web pages?

These problems have promoted efficient research of Web mining. Firstly, we will manage data of Web sites well. Secondly, we will mine interesting contents. Thirdly, we will run after and analyze user usage patterns. Based on the above, it could high effectually provide useful information for Web users and help users to make use of the Web resources efficiently.

II. WEB MINING

Web mining is to explore interesting information and potential patterns from the contents of Web page, the information of accessing the Web, page linkages and resources of e-commerce by using techniques of data mining, which can help people extract knowledge, improve Web sites design, and develop e-commerce better.

The objects of Web mining include: sever logs, Web pages, Web hyperlink structures, on-line market data, and other information.

. Web logs: When people browse Web sever, sever will produce three kinds of log documents: sever logs, error logs, and cookie logs. Through analyzing these log documents we can mine accessing information.

. On-line market data: used as storing e-commerce information in e-commerce sites.

. Web pages: Most of existing web mining methods are used in Web pages of according with HTML standard.

. Web hyperlink structures: the Web pages are all connected by hyperlinks, in which there is very important mining information. So Web hyperlinks are very authoritative resources.

. Other information: main are composed of user registrations that can help mine better.

III. WEB CONTENT MINING

Web content mining contains two parts: searching result mining, and html Web page mining.

A Searching Result Mining

1) Automatic classification in documents using searching engine

Search engine can index a mass of disordered data on Web. For

0-7803-7268-9/01/$10.00 ©2001 IEEE.

example, firstly, Web crawlers down load Web pages from Web sites. Secondly, searching engine extracts describable index information from these Web pages to store them with URL into searching engine base. Thirdly using data mining methods we automatically classify them into usable Web page classification system organized by hyperlink structure.

2) Implement friendly interface from searching results

In the classification system there are many unrelated information. If we can analyze and cluster them in the searching results, searching effect will be high improved.

B Text mining of Web

Text mining is a comprehensive technique. It relates to data mining, computer language, information searching, nature language comprehension, and knowledge management. Text mining uses data mining techniques in text sets to find out connotative knowledge. Its object type is not only structural data but also semi-structural data or non-structural data. The mining results are not only general situation of one text document but also classification and clustering of text sets. The basic frame of text information mining is shown in Fig.1.

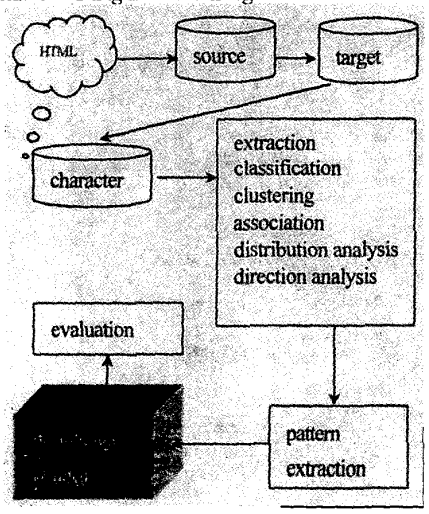

Fig. 1 The frame of text information mining

1) Data selection

Local text base is formed by integrating documents that are waiting for mining and distributed in many Web severs by using information accessing technique.

2) Data preprocessing[3]

We usually extract characteristic metadata as foundation and store them into text character base by using enlightening rules. In order to receive good mining effect we must have neat and exact data and get rid of disordered and redundant data. Firstly, understand users' requests and extract knowledge sources related to transactions from resource. Secondly, clean, transform, and predigest these knowledge sources. Finally produce target data set that is two dimensions table. In the table landscape orientation is metadata and portrait is attribute. After preprocessing the target data set usually has some characters.

. Unify and constrain heterogeneous data.

. wipe off unrelated, noisy and blank data, deal with lost and repeated data.

. Reduce dimensions and promote knowledge discovery effect by using switching, inducing, constraining, and so on.

. Wipe off unrelated attributes to reduce data dimension by using selecting and sampling.

C Collecting construct character , and building character expression

Web text mining is to mine html document sets that are not constrained. So we should transform them into two dimensions table in which data can reflect their characters. We usually adopt TFIDF expression. But there is a very serious problem that the expression will bring quite high vector dimensions. The selection of characters certainly becomes key that directly influences the effect of text mining.

1) Text expression

. Text information is transformed into two dimensions table in which every row is a character set and every column is a character word.

. Classifying words and banning words: Firstly, elect describable character words. Secondly scan the document set many times and wash out words of low frequency. In the end we also eliminate words of high frequency, but vain meaning, such as ah, eh, oh, o.

2) character extraction

Character extraction is only one method that can solve high character vector dimensions brought by text mining. We will give some methods of character extraction.

. Character extraction based on authority-weigh function.

Every character word will obtain an authority-weigh value by calculating authority-weigh function. The higher arisen frequency of character word is, the stronger its ability is in reflecting text theme. Then we will pay it a bigger authority value. In addition, if it

is title, key, and long word it certainly has bigger authority value. Every character word will be sorted. Then we will further select the size of character set. The size must be obtained from experiments.

. Character extraction based on main composition analysis in statistics analysis.

Main idea of the method is to use a comprehensive character word instead of a few primary character words in a describable page to achieve dimensions reduction. In addition, we also use data cube method and induction attribute method to reduce vector dimensions through summing up a lot of data to a higher level.

D Text mining

After collecting, selecting and extracting text set forms character base that will be a foundation of text mining. Based on the above, we can do extraction, classification, clustering, association analysis, distribution analysis, and direction prediction.

1) Text extraction

Text extraction is to extract central meaning that can briefly describe text document in generalization. Then users can understand central meaning, but needn't browse entire text. The method is especially used in searching engine that usually need to give text extraction. Because some searching engines always give the preceding sentences, the best way is to obtain central meaning of text or central text set of text sets by using some algorithms. Based on the above, searching efficiency will be greatly advanced, and will be convenient of users' selecting searching results.

2) Text classification

Firstly, a lot of text documents are quickly and efficiently auto-classified. Secondly, every classified text is placed in a suitable theme range. So it is convenient of users' searching and browsing text documents.

We usually use Navie Bayesian classifier and K-Nearest Neighbor to mine text information. In text classification, firstly, predefine documental classification. Secondly, confirm predefined character through numerating training documental character. In the end, calculate testing documental classification and similarity of predefined documental classification by algorism calculation. Then high similar documents are in same classification. The size of similarity will be measured by prearranged value. If few documental similarities incline to zero, the situation is permitted. If the thing is not like this the selection of predefined classifications is unsuitable. Then we must reselect predefined classifications. In the selection there are usually two phases: training and classifying.

K-Nearest Neighbor Document classification algorism:

. Select predefinition character classification, $Y=\{y_1,\ldots y_i,\ldots y_m\}$

. Locate training document set, $X=\{x_1,\ldots x_j,\ldots x_n\}$ $v(x_j)$ is vector character of x_j

. Every $v(y_i)$ in Y is defined by $v(x_j)$ through training $v(x_j)$ in X.

. Testing document set, $C=\{c_1,\ldots c_k\ldots c_p\}$ c_k in the C is a waiting-classified document, our job is to calculate the similarity between $v(c_k)$ and $v(y_i)$, $\mathrm{sim}(c_k, y_i)$.

. Select document c_k, its similarity to y_i is the largest, then c_k is in the classification of y_i, $\max(\mathrm{sim}(c_k, y_i))$ $i=1,\ldots m$.

Similarity is defined by overlapping degree of two character vectors.

3) Text clustering

Theme classification need not predefined. But we must classify documents into many clusters. In same cluster all documental similarities are requested much higher, or else, are requested much lower. As a rule, related document clusters queried by users are near. So if we use the situation in searching engine results range browsed by users is greatly reduced. In addition, if classified cluster is very big we will reclassify it until users are satisfied at smaller searching range. Arrangement cohesion and level partition are clustering methods in common use.

The process of arrangement cohesion method is as following:

. In defined document set, $W=\{w_1,\ldots w_i,\ldots w_m\}$, every document w_i is a cluster c_i, every cluster set is C that is a clustering $C=\{c_1,\ldots c_j,\ldots c_m\}$.

. For random two clusters c_i and c_j, calculating their similarity $\mathrm{sim}(c_i, c_j)$, If the similarity between c_i and c_j is the largest we will place c_i and c_j in a new cluster. In the end, we will form a new clustering, $C=\{c_1,\ldots c_{m-1}\}$.

. repeat above jobs until leave only one.

The entire process of arrangement cohesion method will form a tree that reflects nesting relationship of similarity in documents. The method has high accuracy. But its pace is very slow due to comparing similarity in all clusters. If document set is bigger the method is not suitable.

The process of level partition method is as following:

. Firstly we should divide document set into clustering seeds (clusters) through optimizing a evaluating function according to certain principle. $R=\{R_1,\ldots R_j,\ldots R_n\}$, n must be predefined.

. For every document in document set W, $W=\{w_1,\ldots w_i,\ldots w_m\}$,

calculate its similarity to seed R_j, $sim(w_i, R_j)$, then select larger similarity document, and fall the document in the R_j cluster.

. repeat above jobs until all documents are fallen in defined clusters.

This method has characters of stable clustering result and quick pace. But we must predefine seeds and its quantity, which directly influences the clustering effect.

4) Text association analysis

Firstly, the relationships of different words are found out from document set. Secondly, through association algorisms we can explore arisen model of words from a lot of documents. In the end, we should further discover text information related to assignments.

5) Distribution analysis and direction prediction

Through analyzing Web documents we can obtain relative distributing circumstance of special data in some historic period, and can predict future development.

E Pattern quality evaluation

Web mining can be regarded as process of machine learning. Result in machine learning is a knowledge pattern. Main part of machine learning is to evaluate produced pattern. We usually classify document sets into training set and testing set. Then to study and test repeatedly in training set and testing set. In the end, average quality is used to evaluate model quality. In machine learning we often use methods of classification accuracy, precision, recall, and information score.

IV. WEB USAGE MINING

Web usage mining is to mine Web log to discover user accessing patterns of Web pages. Through analyzing and exploring regularities in Web log records we can identify customers for e-commerce, enhance the quality of Internet information services to the users, and improve Web server system performance. In addition, Web sites improve themselves by learning from user accessing patterns. Analyzing of Web log may also help build customized Web services for individual user [9].

At present, we often use tools of pattern discovery and pattern analysis. They offer analyses of user actions and filtration of data, and mining knowledge from data collection by using Artificial Intelligence, Data Mining, Psychology and Information Theory. After finding out accessing patterns we often use corresponding analyzing techniques to comprehend, explain and reveal these patterns. For example, On-line analytical processing technique, Data cube predigesting user usage pattern analysis, Sql-like searching knowledge discovered.

A Problems in Web usage mining

Usage mining contains two phrases [8]: In the first phrase, the Web log needs to be cleaned, identified, integrated and transformed. Based on the above, Exploring and analyzing patterns are often needed.

Problems existed:

. Physical structures of Web sites are different from user accessing patterns

. It is very difficult to find out users, sessions and transactions.

B Web usage mining

Web usage mining contains three parts:

. Data preprocessing: identify users, identify accessing operations, perfect paths, identify transactions, integrate data and transform data[5][8]. In this phrase, Web log can be transformed to transaction forms that are fit for data mining in different fields.

. Adopt data mining methods in different fields, such as association, sequential pattern, path analysis, classification and clustering, to discover user usage patterns.

. Model analysis: OLAP, visualization, knowledge searching and intelligent agent.

C Example of Web usage mining

In this example using classification and clustering mining methods, the accessing rules can be got through analyzing users quantity. Then the Web designer can recommend different mass services at varying time according to the rules of users who access Web sites. The good service quality will effectually advance quantity of users in the Web sites. The process is as follows.

. Identify identity of users who access Web sites, analyze special users, find out worthy users through their accessing degree, staying time and loving degree to Web sites.

. Analyze special subjects and Web contents deeply. For example, National Day activity, tour introduction and so on. Realize more mutual relations between users and Web contents. Find out attractive services and commodities for users.

. According to effect of accessing Web sites activity and condition of browsing Web pages, we can plan and evaluate contents of Web sites.

Base on some simulated test data, we express users accessing degree analysis table of a Web site, and analyze recommended services at varying hours during one day as shown in Fig2, Fig.3 and Fig.4.

V. WEB STRUCTURE MINING

WWW is a global information system that was composed of all Web sites. Every one was hyperlinked by many Web pages. The variable hyperlinks contain many high-quality semantic clubs to a page's topic[6]. So it is beneficial to make good use of such semantic information in order to achieve crucial information by analyzing Web linkage analysis.

"What is Web structure mining?" People using mining methods to gain useful knowledge from Web structure, find out important Web pages, and further improve design *plan* for the Web sites.

Time	Users	Time	Users
00:00--00:59	936	12:00--12:59	2466
01:00--01:59	725	13:00--13:59	1432
02:00--02:59	433	14:00--14:59	1649
03:00--03:59	389	15:00--15:59	1537
04:0-0--04:59	149	16:00--16:59	2361
05:00--05;59	118	17:00--17:59	2053
06:00--06:59	126	18:00--18:59	2159
07:00--07;59	235	19:00--19:59	1694
08:00--08:59	399	20:00--20:59	2078
09:00--09:59	1414	21:00--21:59	2120
10:00--10:59	2424	22:00--22:59	1400
11:00--10:59	2846	23:00--23:59	1463

Fig. 2 User statistics table at different time

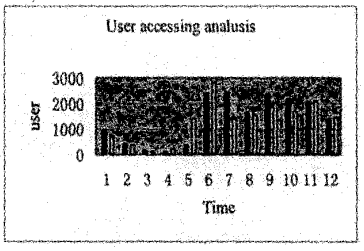

Fig. 3 user accessing analysis

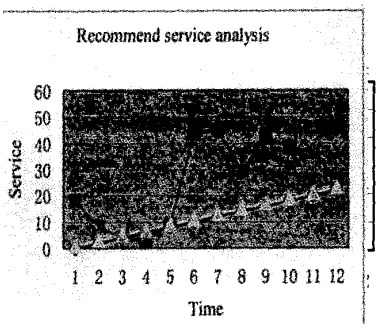

Fig . 4 recommended services analysis

VI. CONCLUSION

Web Mining is a new research field that has a great prospect. And its technology has wide application in the world. Such as text data mining on the Web, time and spatial sequence data mining on the Web, Web mining for the e-commerce system, hyperlink structure mining of Web site and so on. Up to now, the Web mining technology has still been faced with many challenges.

VII. REFERENCE

[1] Fayyad U etal. "The KDD process for extracting useful knowledge from volumes of data", Communications of the ACM, 1996, 39(11)

[2] Hahn U, Schnattinger K. "Deep knowledge discovery from natural language texts", In: Proc of the 3rd Int'l Conference Knowledge Discovery and Data Mining ; Newport Beach, 1997.

[3] Wang Wei qiang, "Text Ming on the Internet Computer Science", 2000

[4] Wang Ji chen, "Research on Web Text Mining", Journal of computer Research&Development, 2000,37(5)

[5] Chen En hong, "Web Usage Mining:Discovering User Behavior Patterns From Web Data", Computer Science, 2001,28(5)

[6] Yang Xiao hua, "Hyperlink Structure Mining of Web Sites", Computer Project&Application, 2001,8

[7] wang shi, "Web Mining", Computer Science, 2000,27(4)

[8] Wang xiao yan, "Web Usage Mining", PH.D thesis 2000,3

[9] Liu Jun, "Web Usage Mining", Master thesis 2000