Web Mining: Knowledge Discovery on the Web

Wang Jicheng*, Huang Yuan*, Wu Gangshan and Zhang Fuyan****
State Key Laboratory for Novell Software Technology
Department of Computer Science and Technology, Nanjing University
Nanjing, Jiangsu 210093, P.R. China
***{hy|wjc}@graphics.nju.edu.cn,**
****{gswu|fyzhang}@netra.nju.edu.cn**

ABSTRACT

With the flood of information on the Web, Web mining is a new research issue which draws great interest from many communities. Currently, there is no agreement about Web mining yet. It needs more discussion among researchers in order to define what it is exactly. Meanwhile, the development of Web mining system will promote its research in turn. In this paper we present a preliminary discussion about Web mining, including the definition, the relationship between information mining and retrieval on the Web, the taxonomy and the function of Web mining. In addition, WebTMS, a prototype of Web text mining system, was designed. WebTMS is a multi-agent system which combines text mining and multi-dimension document analysis in order to help user in mining HTML documents on the Web effectively.

1. INTRODUCTION

Buried in the enormous, heterogeneous and distributed information on the Web was knowledge with great potential value. With the rapid development of the Web, it is urgent and important to provide users with tools for efficient and effective resource discovery and knowledge discovery on the Web. Although the Web search engine assists in resource discovery, it is far from satisfying for its poor precision. Moreover, the target of the Web search engine is only to discover resource on the Web. As far as knowledge discovery is concerned, it is not equal to at all even with high precision. Therefore, the research and development of new technology further than resource discovery is needed.

Data mining is used to identify valid, novel, potentially useful and ultimately understandable pattern from data collection in database community [1]. However, there is little work that deals with unstructured and heterogeneous information on the Web. Web mining is a new research issue under dispute which draws great interest from many communities. Currently, there is no agreement about Web mining yet. It needs more discussion among researchers in order to define what it is exactly. Meanwhile, the development of Web mining system will promote its research in turn.

In this paper a preliminary discussion about Web mining is given. Firstly we present the definition of Web mining and expound the relationships between Web mining, traditional data mining and information retrieval on the Web. Then we present the taxonomy of Web mining and briefly describe the functions of Web text mining. In addition, a prototype of Web text mining system WebTMS is analyzed. WebTMS is a multi-agent system we are developing, which combines text mining and multi-dimension document analysis technologies in order to help user in mining HTML documents on the Web effectively. Finally, we present the conclusions and challenges for Web mining in future.

2. WEB MINING AND WEB INFORMATION RETRIEVAL

2.1 The definition of Web mining

Web mining is an integrated technology in which several research fields are involved, such as data mining, computational linguistics, statistics, informatics and so on. Different researchers from different communities disagree with each other on what Web mining is exactly. Many unrelated projects have explored different aspects of this problem as well. We present a more general definition of Web mining as follows.

Definition 1. Web mining is the activity of identifying patterns p implied in large document collection C, which can be denoted by a mapping $\xi : C \rightarrow p$.

Since Web mining derives from data mining, its definition is similar to the well-known definition of data mining [1]. Nevertheless, Web mining has many unique characteristics compared with data mining. Firstly, the source of Web mining is web documents. We consider the use of the Web as a middleware in mining database and the mining of logs, user profiles on the Web server still belong to the category of traditional data mining. Secondly, the Web is a directed-graph consists of document nodes and hyperlinks. Therefore, the pattern identified can be possibly about the content of documents or about the structure of the Web. Moreover, the Web documents are semi-structural or non-structural with little machine-readable semantic while the source of data mining is confined to the structural data in database. As a result, some traditional data mining methods are not applicable to Web mining. Even if applicable, they must be based on the preprocessing of documents.

2.2 Web information retrieval

Definition 2. Web information retrieval is the process to find a subset S of appropriate number of documents relevant to a

0-7803-5731-0/99$10.00 ©1999 IEEE

certain query q from large document collection C, which can also be denoted by a mapping $\zeta:(C,q)\rightarrow S$.

Since 1960, there have been many achievements in the field of information retrieval, such as index model, document representation and similarity measure. These achievements were applied on the Web successfully, which gave rise to search engines. In recent years, some researchers applied database concept to the Web and presented some new methods of modeling and querying the Web at a finer granularity level than pages, such as WebOQL [2], Lorel [3], etc. These methods can retrieve not only the hyperlink between Web pages but also the internal structure within a web page.

Web information retrieval and Web mining have different goals. Although Web mining is further than Web information retrieval, it does not intend to replace Web information retrieval. Instead they are two technologies supplement each other. On the one hand, each has its strong points and applications in point. On the other hand, Web mining can be utilized to increase the precision of information retrieval and improve the organization of retrieval results as well, which will bring the information retrieval system into the next generation.

3. A TAXONOMY OF WEB MINING

The diversity of information on the Web leads to the variety of Web mining, as shown in figure 1. According to the type of source, Web mining can be roughly divided into two domains: Web content mining and Web structure mining. The former is the process of extracting knowledge from the content of Web documents, while the latter is the process of inferring knowledge from the organization and links on the Web.

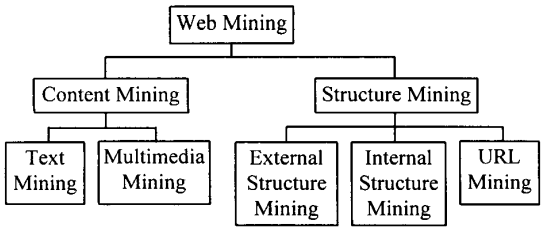


Figure 1 A taxonomy of Web mining

Web structure mining can be further divided into external structure (hyperlink between web page) mining, internal structure (of a web page) mining and URL mining. Craven et al. [4] used first-order learning in categorizing hyperlinks to estimate the relationship between web pages. They also made use of the anchor text in hyperlink to categorize the target web page. Brin et al. [5] took both the citation counting of referee pages and the importance of referrer pages into account to find pages that are "authorities" on particular topics. Spertus et al. [6] proposed some heuristic rules by investigating the internal structure and the URL of web pages.

Web content mining can be divided into text mining (including text file, HTML document, etc.) multimedia mining. Although multimedia mining is drawing more and more interests, text mining is the most fundamental and important task as text is the primary information vehicle. In this paper, we only discuss text mining on the Web. As to multimedia mining on the Web, the readers interested in can refer to [7], in which a prototype of Web multimedia mining system is introduced.

The main categories of Web text mining are text categorization, text clustering, association analysis, trend prediction and so on.

a) Text categorization: Given a predefined taxonomy, each document in collection C is categorized into one appropriate class or more. In this way, it is not only convenient for users to browse documents but also easier to search documents by specifying class. Currently there are many text categorization algorithms, in which the most commonly used are k-Nearest Neighbor algorithm [8], Naive Bayes algorithm [9], etc.

b) Text clustering: The difference between text clustering and categorization lies in that no taxonomy is predefined for clustering. The goal of text clustering is usually to divide documents collection C into a set of clusters such that inter-cluster similarity is minimized and intra-cluster similarity is maximized. We can apply text clustering to the organization of retrieval results returned by search engine. Then users merely need to examine the clusters relevant to their queries, which dramatically reduces the time and the efforts spent in sifting through the long list of documents. Numerous text clustering algorithms have been proposed so far, all of which fall into two types. One is hierarchical clustering such as G-HAC algorithm [10], the other is partitional clustering represented by k-Means algorithm [11].

c) Association analysis: Extraction of the relationship between phases and words in documents. For example, Wang et al. [12] tested their algorithm on the Web Movie Database and found several patterns about director, cast, writer and so on.

d) Trend prediction: Prediction of the value of given data at the specific time in future. For example, Wüthrich et al. [13] predicted stock markets using information contained in articles published on the Web and obtained good results.

Note that the methods of Web text mining are similar to mining of flat text files to some extent. However, additional information is conveyed by the tags in Web documents, such as <Title>, <Heading> and so on, which can be exploited to increase the quality of Web text mining.

4. A WEB TEXT MINING SYSTEM PROTOTYPE WEBTMS

WebTMS is a Web text mining system prototype we are developing, which adopts multi-agents architecture and combines text mining and multi-dimension document analysis in order to help user in mining HTML documents on the Web effectively, as shown in figure 2.

4.1 Multi-dimension document analysis and mining

Both information retrieval systems and text mining systems require that resource and knowledge be visualized in an intuitionistic manner convenient for users to explore. Traditional search engines always return long ordered lists of Web documents to users, which cause a great trouble for them to

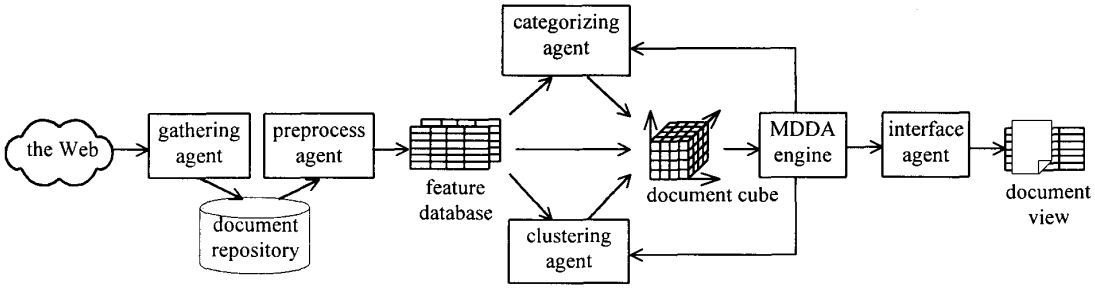


Figure 2 Web text mining system prototype WebTMS

examine the results. These problems are worsened when the number of documents returned is large. It is shown by the research we did on users' behavior that users often want to explore document resource from various viewpoints, including the properties of documents and the relationship among documents, much more than a simple fixed-rank list. OLAP tools are powerful in data warehouse, which provide users multi-dimension views of data [14]. Although there are intrinsical differences between data warehouse and Web documents collection, we believe that Web documents analysis and mining can still benefit from OLAP technology. Drawing on the experience of multi-dimension data analysis, we introduce the document cube and multi-dimension document analysis engine.

Definition 3. Dimension d is the viewpoint from which users explore the documents. For instance, sometimes users want to examine the documents according to date or author. Users are also interested in the documents of certain subject. In these cases, such metadata about documents as date, author and subject are all dimensions. Metadata can be categorized into two types: descriptive ones and semantic ones. Descriptive metadata include title, date, size, document type etc, while semantic metadata include author, organization, subject and so on.

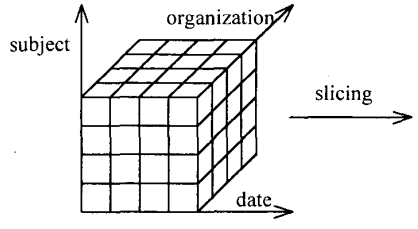
Definition 4. Document cube $C_{Document}$ is the super cube $(d_1, d_2, \cdots, d_m, Document)$, where $Document$ is the core around which metadata (dimension d_i) circles.

Based on document cube, multi-dimension document analysis engine can apply various analytic operations such as slicing, dicing, rotation, drilling-down and rolling-up to create a variety of document views. Thus users can look into document resources from multiple viewpoints and comprehend the connotation embodied in. For instance, users can slice the cube to create a view of documents published by a certain organization. In this view, users can further classify documents according to the subject, then rank the documents of the same subject by date. Users not only can observe the overall characteristics of each subject by rolling up to fold the view, but also can look into concrete characteristics of each document by drilling down to unfold the view. As shown in figure 3.

In addition, multi-dimension document analysis engine has the statistic function to show metadata distribution over document collection. For example, by comparing the documents on a given subject published by different organizations for certain duration, we can answer some questions which can not be handled by traditional search engines, such as "which university in China published most papers on video conference in last five years?"

Note that the above-mentioned document cube and multi-dimension document analysis technology are based on the pre-process of documents and depends on text mining technology. Some dimensions of document cube, such as date and author, come from the metadata obtained by document pre-processing, while other dimensions such as subject are created by text clustering or text categorization. At the same time, multi-dimension document analysis engine facilitates text mining with an effective visualization method and selection means including document selection and feature pruning. By document selection users can eliminate noise documents to improve the quality of mining or confine the input source of mining to a subset of documents collection. By feature pruning users can filter out those dimensions useless for mining to reduce the number of dimensions. For example, the date dimension can be left out when documents are classified with respect to the content. The results of selection can be presented to users or serve as input source of text mining, as shown in figure 2.

4.2 System components



organization="Dept. Of CS, Nanjing Univ."

		Date →		
subject ↓		1990	...	1998
multimedia	graphics	7		11
	...			
	video conference	0		5
...	...			
operating system	...			

Figure 3 Document cube and multi-dimension document view

a) Document Gathering Agent (DGA): DGA gathers the documents to be mined which may distribute over several Web

servers and store them in the document repository of WebTMS. Users can specify a list of initial URLs to avoid random gathering. A more flexible measure can allow users to put forward gathering strategies, such as subject or network domain. The strategy will be stored in the profile and a list of URLs will be created automatically by the URL manager of DGA, as shown in figure 4.

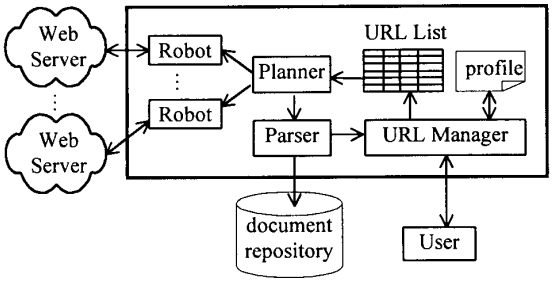


Figure 4 Document Gathering Agent

b) Document Preprocessing Agent (DPA): Using heuristic rules and NLP technology to extract metadata and representation of documents, as shown in figure 5. The vector-space model (VSM) [15], which has been used widely and shown good effect in last decades, is employed by DPA. Each document d is represented by a vector $\overrightarrow{V_d} = (t_1, w_{d1}; \cdots; t_i, w_{di}; \cdots; t_n, w_{dn})$ or $\overrightarrow{V_d} = (w_{d1}, \cdots, w_{dn})$ compactly, where t_i is a word (or a term) in document collection and w_{di} is the weight of t_i in d. In DPA, w_{di} is defined as $w_{di} = tf_{di} \times \log(N/idf_{di} + 0.5)$, where tf_{di} is the occurrences of t_i in d, N is the number of total documents in the collection and idf_{di} is the number of documents in which t_i appears. After normalizing, $w_{di} = \frac{tf_{di} \times \log(N/idf_{di} + 0.5)}{\sqrt{\sum_{i=1}^{n} \left(tf_{di}^{\ 2} \times \log^2(N/idf_{di} + 0.5)\right)}}$. The matadata and representation of documents are all stored as structured data in the feature database for future use in text mining.

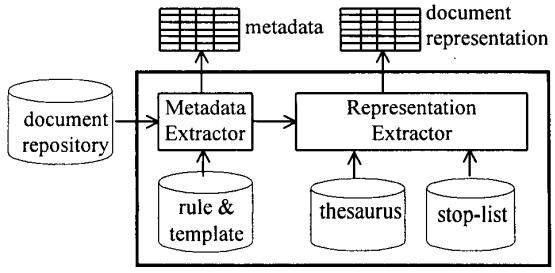


Figure 5 Document Preprocessing Agent

c) Document Categorization Agent (DCAA): categorizing the documents of a collection (or a subset) in term of a predefined taxonomy, as shown in figure 6. Currently, VSM-based algorithm is adopted and the similarity between two documents is defined as $sim(d,c) = \frac{\sum_{i=1}^{n} (w_{di} \times w_{ci})}{\sqrt{\sum_{i=1}^{n} w_{di}^{\ 2}} \times \sqrt{\sum_{i=1}^{n} w_{ci}^{\ 2}}}$. We will equip DCAA with new categorization algorithms or the combination of several ones to improve the performance.

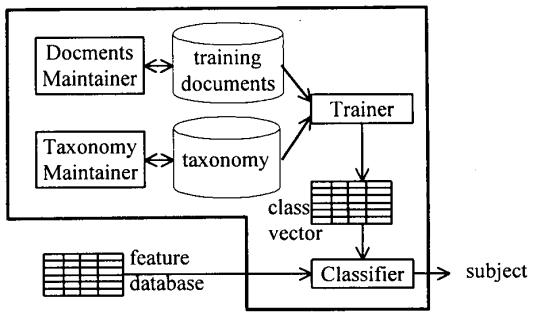


Figure 6. Document Categorization Agent

d) Document Clustering Agent (DCLA): DCLA can group documents into several clusters, as shown in figure 7. Considering the large amount of documents, the clustering methods ought to be fast enough to make online clustering possible. The method adopted by DCLA is k-Means algorithm [11] which has been proved much faster than hierarchical ones in many experiments. DCLA divides documents into several clusters and creates summarization for each cluster to give users a clear sense of the topics.

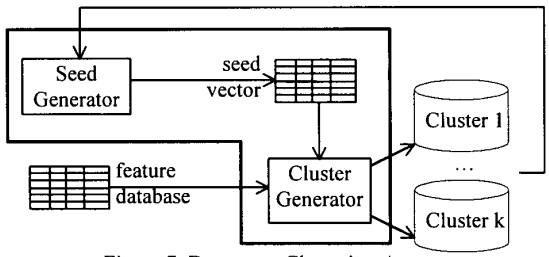


Figure 7. Document Clustering Agent

e) Multi-Dimension Document Analysis Engine (MDDAE): WebTMS introduces document cube and multi-dimension document analysis technology to provide multi-dimension views of documents for users. MDDAE also has the function of statistical analysis to reveal the distribution or trend in documents. Moreover, it provides selection means to facilitate the operation of Web text mining components such as DCAA, DCLA and so on.

f) User Interface Agent (UIA): As the bridge between users and MDDAE, UIA provides visual interface for users. It can translate users' requests into specialized languages and pass them to MDDAE, DCAA, DCLA, etc., then shows the multi-dimension document views and documents returned to users.

Each agent, as a component of the system, can accomplish relatively autonomous work. These components can be located in a computer or several computers distributed over the net. Moreover, new components are easy to be added into WebTMS due to the high modularization. At the same time, the output of one component can be imported into another component so those components can be assembled in a Producer/Consumer chain. Agents on the chain cooperates with each other to accomplish the whole process of Web text mining. Thus, users can construct a complicated scheme for Web text mining by combining components appropriately.

4.3 System behavior

WebTMS supports the complete cycle of Web text mining. Users can perform document analysis and mining with several iterations until satisfying results is obtained.

Firstly, users present some gathering strategies for DGA to gather Web documents. Secondly, DPA extracts metadata and representation of documents. Then users have many choices to make, including:

a) Using MDDAE to get multi-dimension document views, each of which corresponds with a subset of documents.

b) Using DCAA to classify document collection or a subset in term of a predefined taxonomy,

c) When the predefined taxonomy is not consistent with the inherent topics of the document collection (or subset), DCAA can not classify documents correctly. Here, users can modify the taxonomy and update training documents.

d) In the above case, an alternative is to evoke DCLA to group documents into clusters. Since each cluster is a subset of documents, if the cluster which users are interested in has too many documents, it can be further divided into several sub-clusters.

5. CONCLUSIONS

With the information overload, Web mining is a new and promising research issue to help users in gaining insight into overwhelming information on the Web. Workshops on Web mining have been already or will be held to discuss its principle, architecture and algorithm in several international conferences, such as KDD'97, KDD'99, etc. In this paper, we present a preliminary discussion about Web mining, including the definition, the taxonomy and the function, and introduce a Web text mining system prototype WebTMS we are developing. There still remain many areas for further research, such as the design of efficient algorithms for very large document collections, the use of XML specifications [16] to describe and extract metadata about Web documents, the development of more components to enrich the function of WebTMS, and so on.

6. REFERENCES

[1] Usama Fayyad et al, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, Vol. 39, No. 11, Nov. 1996, pp. 27-34.

[2] Gustavo Arocena and Alberto Mendelzon, "WebOQL: Restructuring Documents, Databases and Webs", In Proceedings of International Conference on Data Engineering (ICDE), Orlando, Florida, 1998.

[3] Serge Abiteboul, Dallan Quass, Jason McHugh,Jennifer Widom, and Janet Wiener, "The Lorel Query Language for Semistructured Data", International Journal on Digital Libraries, Vol. 1, No. 1, 1997, pp. 68-88.

[4] M. Craven, S. Slattery and K. Nigam, "First-Order Learning for Web Mining", In Proceedings of the 10th European Conference on Machine Learning, Chemnitz, 1998.

[5] Sergey Brin, Lawrence Page, "The Anatomy of Large-scale Hypertextual Web Search Engine", In Proceedings of the Seventh International World Wide Web Conference, 1998.

[6] Ellen Spertus, "ParaSite: Mining Structural Information on the Web", In proceedings of the Sixth International World Wide Web Conference, 1997.

[7] Osmar R. Zaïane, Jiawei Han, Ze-Nian Li, Sonny H. Chee and Jenny. Y. Chiang, "MultiMedia-Miner: A System Prototype for MultiMedia Data Mining", In Proceedings of 1998 ACM-SIGMOD Conference on Management of Data, Seattle, 1998.

[8] Y. Yang, "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval", In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94) 1994.

[9] Tom Mitchell, Machine Learning, McGraw-Hill, 1996.

[10] P. Willet, "Recent Trends in Hierarchical Document Clustering: a Critical Review", Information Processing and Management, Vol. 24, 1988, pp. 577-597.

[11] J. J. Rocchio, "Document Retrieval Systems – Optimization and Evaluation", Ph.D. Thesis, Harvard University, 1966.

[12] Ke Wang and Huiqing Liu, "Schema Discovery from Semi-structured Data". In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Newport Beach, 1997.

[13] B. Wüthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang and W. Lam, "Daily Prediction of Major Stock Indices from Textual WWW Data", In Proceedings of the 4th International Conference on Knowledge Discovery, New York, 1998.

[14] S. Chaudhuri and U. Dayal, "An Overview of Data Warehousing and OLAP Technology", ACM SIGMOD Record, Vol. 26, 1997, pp. 65-74.

[15] G. Salton, A.Wong and C.S.Yang, "A Vector Space Model for Automatic Indexing", Communications of the ACM, Vol. 18, 1975, pp. 613-620.

[16] Tim Bray, Jean Paoli and C. M. Sperberg-McQueen, "Extensible Markup Language (XML) 1.0 Specification", World Wide Web Consortium Recommendation, 1998.