WEB CONTENT MINING-A STUDY

M.Vanathi
Assistant Professor,
Department of Computer Science
Sri Krishna Engineering College, Chennai.

ABSTRACT

Data mining is accumulating the exact information needed by the user through several steps. Web is huge collection of potential information. Web mining is part of Data Mining where the user find his or her information in the Web. There are three types of Web Mining namely Web Content mining, Web Structure mining and Web Usage mining. This paper focuses on Web Content mining especially the techniques available for Web Content mining.

KEYWORDS:
Web content mining, NLP, Information retrieval

1. INTRODUCTION

The World Wide Web serves as a huge, widely distributed, global information service center for news, education, e-commerce, and many other information services. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. This area of research is so huge today partly due to the interests of various research communities, the tremendous growth of information sources available on the Web and the recent interest in e-commerce. The following sections give overview about Web Content Mining and its techniques.

2. WEB CONTENT MINING

Web content mining is the process to discover useful information from text, image, audio or video data in the web. Web content mining sometimes is called web text mining, because the text content is the most widely researched area. Text mining and web mining are two interrelated fields that have received a lot of attention in recent years. Text mining is concerned with the analysis of very large document collections and the extraction of hidden knowledge from text-based data. Web mining refers to the analysis and mining of all web-related data, including web content, hyperlink structure, and web access statistics. Among the three aspects of web mining, text mining is most closely related to web content mining. However, whereas text mining deals with text documents in general, such as emails, letters, reports, and articles, which exist in intranet and internet environment, web content mining is primarily concerned with the materials on the web only. Web content mining is further divided into Web page content mining and search results mining. The first is traditional searching of Web pages via content, while the second is a further search of pages found from a previous search. The technologies that are normally used in web content mining are NLP (Natural Language Processing) and IR (Information Retrieval).

3. NATURAL LANGUAGE PROCESSING (NLP)

Natural language processing (NLP) is a field of computer science concerned with the interactions between computers and human (natural) languages. The goal of NLP evaluation is to measure one or more qualities of an algorithm or a system, in order to determine whether (or to what extent) the system answers the goals of its designers, or meets the needs of its users. Research in NLP evaluation has received considerable attention, because the definition of proper evaluation criteria is one way to specify precisely an NLP problem, going thus beyond the vagueness of tasks defined only as language understanding or language generation. A precise set of evaluation criteria, which includes mainly evaluation data and evaluation metrics, enables several teams to compare their solutions to a given NLP problem.

Natural language generation systems convert information from computer databases into readable human language. Natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate. Many problems within NLP apply to both generation and understanding; for example, a computer must be able to model morphology (the structure of words) in order to understand an English sentence, but a model of morphology is also needed for producing a grammatically correct English sentence.

NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of artificial intelligence. The term natural language is used to distinguish human languages (such as Spanish, Swahili or Swedish) from formal or computer languages (such as C++, Java or LISP). Although NLP may encompass both text and speech, work on speech processing has evolved into a separate field.

4. INFORMATION RETRIEVAL (IR)

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). It is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the World Wide Web. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis and technologies. IR is interdisciplinary, based on computer science, Mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics and physics. Automated information retrieval systems are used to reduce what has been called "information overload". The following section discusses some of the retrieval methods of text.

4.1 Indexing

The main method in retrieving information from Web is Indexing. The index contains selected term, with principal terms where the terms occurs. Indexing is act of assigning index terms to a documents either manually or automatically. Indexing has three primary purpose in information retrieval.

- To permit easy location of document by topic
- To define topic areas, and hence relate one document to another and
- To predict relevance of a given document to a specified information need.

4.2 Matrix Representations

A matrix is rectangular array of cells holding information. A term-term matrix T , is a square matrix whose rows and columns each represent the vocabulary terms. For this matrix, a nonzero value in cell T_{ij} means that the ith and jth term occurs together in some document or have some other defined relationship.

A document-document matrix, D is a matrix whose rows and columns represent documents. In this matrix, a nonzero value in cell Dij would indicate that the documents have some terms in common or have some other defined relationship, such as an author in common.

4.3 Stop Lists

The stop lists play an important role in retrieving the text. Here the common words such as a, an ,the etc that occur in word are not considered. These common words have two impacts on information retrieval system.

1. The very high frequency word tend to diminish the impact of frequency differences among less common words.
2. These word carry very little meaning by themselves, they may result in a large amount of unproductive processing if left in the text.

4.4 Stemming

Stemming is the concept of removing the word endings. For example, computer, computers, computing, computationally and various other words have the same basic form and all deal with set of closely related concepts. One way to solve this problem is to introduce stemming algorithm, which strips off word endings, reducing them to a common core or stem .For the above example the stem might be comput. For a given document this will bring together the various forms of the word, resulting in a higher frequency count and thus in greater significance for the term.

4.5 Thesauri

The final major problem to be considered here is use of similar or related terms. These cannot be handled by a simple algorithm the variant word forms are, since frequently the words are quite distinct. For example, one may post a letter or mail a letter, with exactly same result. Thesauri are used to address this problem. A good thesaurus may contain both synonyms and antonyms for each word together with broader and narrower terms, and closely related terms. A thesaurus can be used during the query process to broaden a query and ensure that relevant documents are not missed because of a narrow query vocabulary.

5 Conclusion

This paper gives an overview of the Web content mining.The Web presents new challenges to the traditional data mining algorithms that work on flat data. An interesting direction of Web content mining is the recent interest in information integration, which could be in the form of a Web knowledge base or Web warehouse, or in the form of a mediator .At least this is the area where database and other research communities such as IR, AI, and machine learning met recently.How the Web extracts information by using information retrieval techniques. The types of text retrieval are presented. It can be further expanded to User profiles and their use.

7. REFERENCES

[1] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In Proceedings of the 21st annual international ACM SI-
GIR conference on Research and development in information retrieval, pages 104–111, 1998.

[2] J. Borges and M. Levene. Mining association rules in hypertext databases. In proceedings of the Fourth International Conference on Knowledge Discovery and
Data Mining (KDD-98), 1998.

[3] J. Borges and M. Levene. Data mining of user navigation patterns. In Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, pages 31–36, 1999.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In Seventh International World Wide Web Conference, 1998.

[5] A. Buchner, M. Baumgarten, S. Anand, M. Mulvenna, and J. Hughes. Navigation pattern discovery from internet data. In Proceedings of the WEBKDD'99 Workshop
on Web Usage Analysis and User Profiling, 1999.

[6] J. Carbonell, M. Craven, S. Fienberg, T. Mitchell, and Y. Yang. Report on the conald workshop on learning from text and the web. In CONALD Workshop on Learning from Text and the Web, 1998.

[7] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Mining the link structure of the world wide web. IEEE Computer, 32(8):60–67, 1999.

[8] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The tsimmis project: Integration of heterogeneous information
sources. In Proceedings of the 10th Meeting of the Information Processing Society of Japan, pages 7–18, 1994.

[9] W. W. Cohen. What can we learn from the web? In Proceedings of the Sixteenth International Conference on Machine Learning (ICML'99), pages 515–521,
1999.

[10] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1(1), 1999.

[11] O. Etzioni. The world wide web: Quagmire or gold mine. Communications of the ACM, 39(11):65–68, 1996.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In Advances in Knowledge Discovery and Data Mining,
pages 1–34. AAAI Press, 1996.
[13] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: toward a unifying framework. In Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining, pages 82–88, 1996.
[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In Proc. of ACM-SIAM Symposium on Discrete Algorithms, pages 668–677, 1998.
[15] R. Kosala and H. Blockeel. Web mining research: A survey. SIGKDD Explorations, vol. 2, no. 1, pp. 1–15, 2000.
[16] R. Korfhage .Information Storage and Retrieval.Wiley India Edition.