ISSN:0975-9646

An Overview on Data Mining

Vishal[1], Dr Saurabh Gupta[2]

[1] *Cyber Q Consulting Pvt. Ltd.,*
622,DLF Tower A, Jasola, New Delhi-110025 India

[2]*Department of IT, National Informatics Centre*
(Ministry of Communication &IT), Shimla-171001, India

Abstract- In this paper, we present an overview of research issues in Data mining. We discuss mining with respect to web data referred here as web data mining. In particular, our focus is on web data mining research in context of our web Data Mining project called 'A Study on Web Data Mining'. We have categorized web data mining into three areas; web content mining, web structure mining and web usage mining. We have highlighted and discussed various research issues involved in each of these web data mining category. We believe that web data mining will be the topic of exploratory research in near future.

INTRODUCTION

The advent of the World Wide Web has caused a dramatic increase in the usage of the Internet. The World Wide Web is a broadcast medium where a wide range of information can be obtained at a low cost. Information on the WWW is important not only to individual users, but also to the business organizations especially when the critical decision-making is concerned. Most users obtain WWW information using a combination of search engines and browser; however, these two types of retrieval mechanisms do not necessarily address all of a user's information needs. This is particularly true in the case of business organizations that currently lack suitable tools to systematically harness strategic information from the web and analyze these data to discover useful knowledge to support decision making. A recent study provides a comprehensive and comparative evaluation of the most popular search engines. A more recent survey of web query processing has appeared.

The resulting growth in on-line information combined with the almost unstructured web data necessitates the development of powerful yet computationally efficient web data mining tools. Web data mining can be defined as the discovery and analysis of useful information from the WWW data. Web involves three types of data; data on the WWW, the web log data regarding the users who browsed the web pages and the web structure data.

WHY DATA MINING?

The Web—an immense and dynamic collection of pages that includes countless hyperlinks and huge volumes of access and usage information—provides a rich and unprecedented data mining source. However, the Web also poses several challenges to effective resource and knowledge discovery:

- Web page complexity far exceeds the complexity of any traditional text document collection. Although the Web functions as a huge digital library, the pages themselves lack a uniform structure and contain far more authoring style and content variations than any set of books or traditional text-based documents. Moreover, the tremendous number of documents in this digital library has not been indexed, which makes searching the data it contains extremely difficult.
- The Web constitutes a highly dynamic information source. Not only does the Web continue to grow rapidly, the information it holds also receives constant updates. News, stock market, service center, and corporate sites revise their Web pages regularly. Linkage information and access records also undergo frequent updates.
- The Web serves a broad spectrum of user communities. The Internet's rapidly expanding user community connects millions of workstations. These users have markedly different backgrounds, interests, and usage purposes. Many lack good knowledge of the information network's structure, are unaware of a particular search's heavy cost, frequently get lost within the Web's ocean of information, and can chafe at the many access hops and lengthy waits required to retrieve search results.
- Only a small portion of the Web's pages contain truly relevant or useful information. A given user generally focuses on only a tiny portion of the Web, dismissing the rest as uninteresting data that serves only to swamp the desired search results. How can a search identify that portion of the Web that is truly relevant to one user's interests? How can a search find high-quality Web pages on a specified topic?

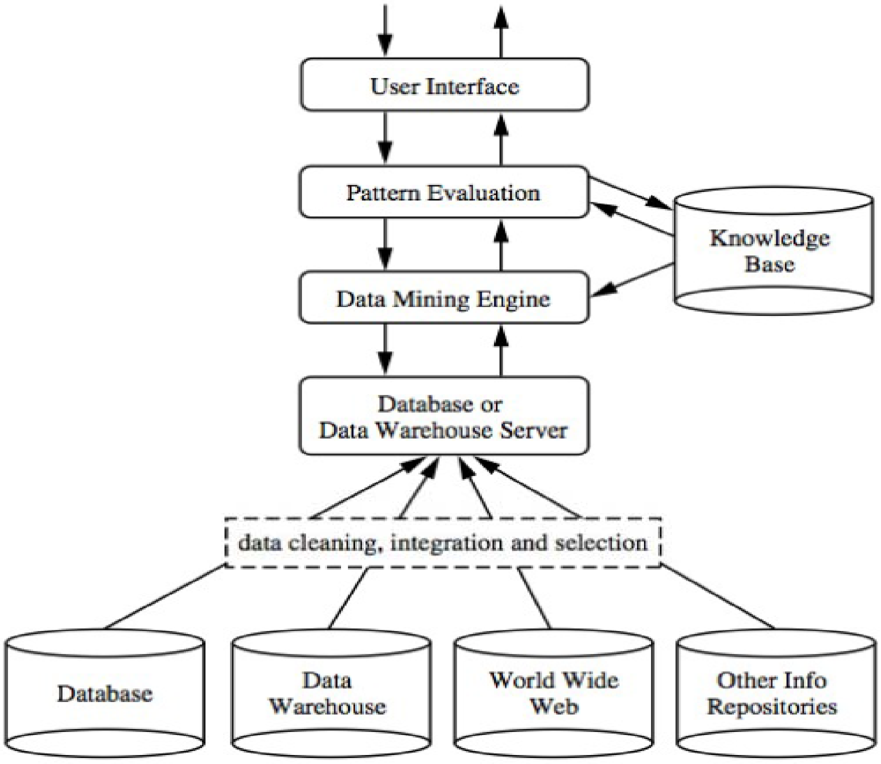
DATA MINING

Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data.

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).

The key properties of data mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases



DATA MINING AND STATISTICS

There is a great deal of overlap between data mining and statistics. In fact most of the techniques used in data mining can be placed in a statistical framework. However, data mining techniques are not the same as traditional statistical techniques.

Traditional statistical methods, in general, require a great deal of user interaction in order to validate the correctness of a model. As a result, statistical methods can be difficult to automate. Moreover, statistical methods typically do not scale well to very large data sets. Statistical methods rely on testing hypotheses or finding correlations based on smaller, representative samples of a larger population.

Data mining methods are suitable for large data sets and can be more readily automated. In fact, data mining algorithms often require large data sets for the creation of quality models.

DATA MINING AND OLAP

On-Line Analytical Processing (OLAP) can be defined as fast analysis of shared multidimensional data. OLAP and data mining are different but complementary activities.

OLAP supports activities such as data summarization, cost allocation, time series analysis, and what-if analysis. However, most OLAP systems do not have inductive inference capabilities beyond the support for time-series forecast. Inductive inference, the process of reaching a general conclusion from specific examples, is a characteristic of data mining. Inductive inference is also known as computational learning.

OLAP systems provide a multidimensional view of the data, including full support for hierarchies. This view of the data is a natural way to analyze businesses and organizations. Data mining, on the other hand, usually does not have a concept of dimensions and hierarchies.

Data mining and OLAP can be integrated in a number of ways. For example, data mining can be used to select the dimensions for a cube, create new values for a dimension, or create new measures for a cube. OLAP can be used to analyze data mining results at different levels of granularity.

Data Mining can help you construct more interesting and useful cubes. For example, the results of predictive data mining could be added as custom measures to a cube. Such measures might provide information such as "likely to default" or "likely to buy" for each customer. OLAP processing could then aggregate and summarize the probabilities.

DATA MINING AND DATA WAREHOUSING

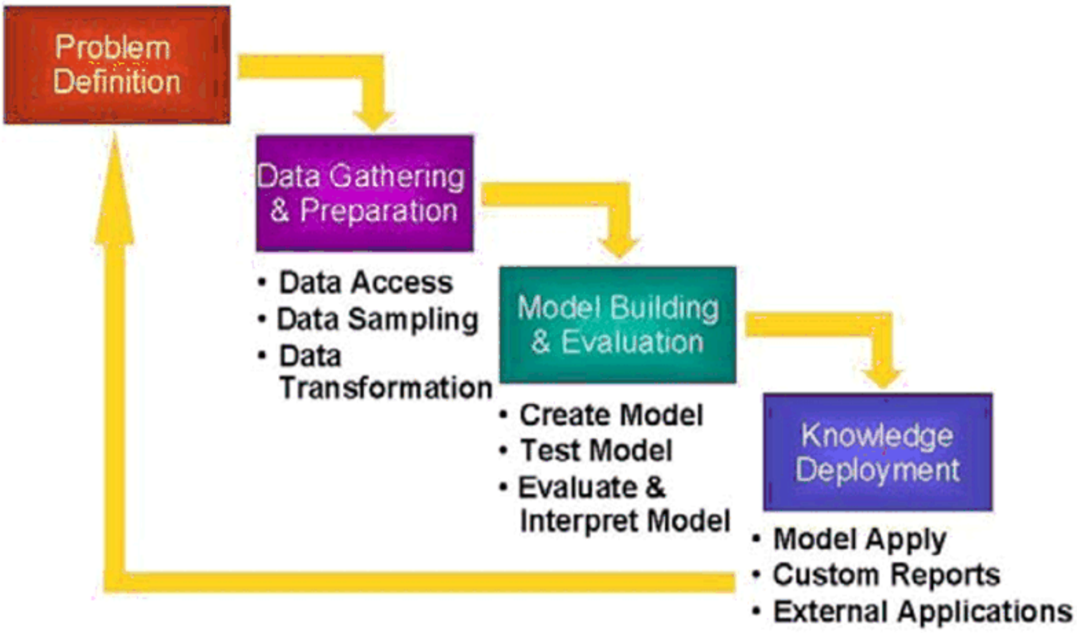
Data can be mined whether it is stored in flat files, spreadsheets, database tables, or some other storage format. The important criteria for the data are not the storage format, but its applicability to the problem to be solved.

Proper data cleansing and preparation are very important for data mining, and a data warehouse can facilitate these activities. However, a data warehouse will be of no use if it does not contain the data you need to solve your problem.

Oracle Data Mining requires that the data be presented as a case table in single-record case format. All the data for each record (case) must be contained within a row. Most typically, the case table is a view that presents the data in the required format for mining.

THE DATA MINING PROCESS

The process flow shows that a data mining project does not stop when a particular solution is deployed. The results of data mining trigger new business questions, which in turn can be used to develop more focused models.



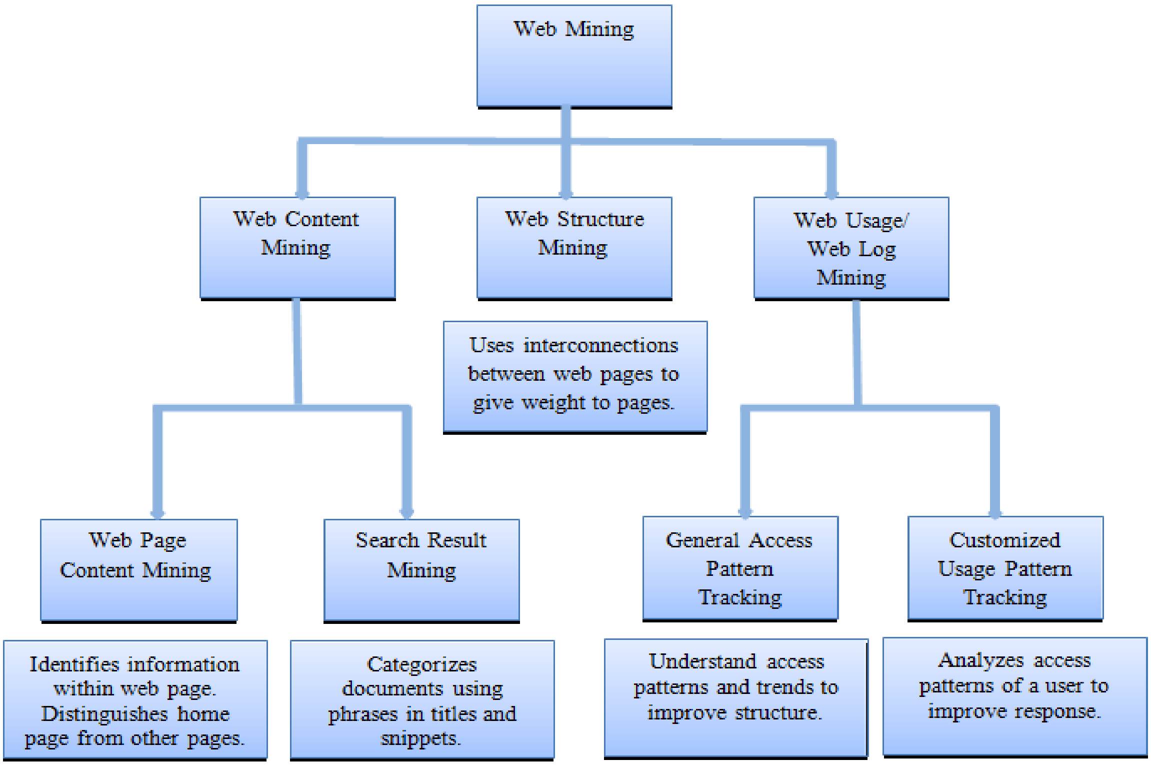

Web Mining
Web Content Mining
Web Structure Mining
Web Usage/ Web Log Mining
Uses interconnections between web pages to give weight to pages.
Web Page Content Mining
Search Result Mining
General Access Pattern Tracking
Customized Usage Pattern Tracking
Identifies information within web page. Distinguishes home page from other pages.
Categorizes documents using phrases in titles and snippets.
Understand access patterns and trends to improve structure.
Analyzes access patterns of a user to improve response.

Fig1: Web Mining Taxonomy

TYPE OF DATA MINING

Web Content Mining -Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited.

Web Structure Mining -The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used. *Hyperlinks* A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an *intra-document hyperlink*, and a hyperlink that connects two different pages is called an *inter-document hyperlink*.

Web Usage Mining - usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications .Usage data captures the identity or origin of web users along with their browsing behavior at a web site. web usage mining itself can be classified further depending on the kind of usage data considered: *Web Server Data* User logs are collected by the web server and typically include IP address, page reference and access time. *Application Server Data* Commercial application servers such as Web logic, 1,2 StoryServer,3 have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

CONCLUSIONS

In this paper, we have discussed some web data mining research issues in context of the ‘A Study On Web Data Mining Project’ at Shri Venkateshwara University, Gajraula, Amroha (Uttar Pradesh). We have defined three types of web data mining. In particular, we discussed web data mining with respect to web structure, web content and web usage. An important part of our project is to design for web data mining to generate some useful knowledge from the WWW data. Currently we are exploring the ideas discussed in this paper.

REFERENCES

1. H. Vernon Leighton and J. Srivastava. Precision Among WWW Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos. http://www.winona.msus.edu/is-f/libraryf/webind2/webind2.htm, 1997.
2. R. Cooley, B. Mobasher and J. Srivsatava. Web Mining: Information and Pattern Discovery on the Word Wide Web. Technical Report TR 97-027, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.
3. J. Han, Yue Huang, et al. Intelligent Query Answering by Knowledge Discovery Techniques, IEEE TKDE, 1996.
4. S. K. Madria, M. Mohnia, J. Roddick. Query Processing in Mobile Databases Using Concept Hierarchy and Summary Database. In proceedings of 5th International Conference on Foundation of Data Organization, Japan, Nov. 1998. 16
5. Sourav S. Bhowmick, S. K. Madria, W.-K. Ng, E.-P. Lim, Web Bags : Are They Useful in Web warehouse? In proceedings for 5th International Conference on Foundation of Data Organization, Japan, Nov. 1998.
6. T. Bray, Measuring the Web. In Proceedings of the 5th Intl. WWW Conference, Paris, France, 1996.
7. Wee-Keong Ng, Ee-Peng Lim, Chee-Thong Huang, Sourav Bhowmick, Fengqiong Qin. Web Warehousing : An Algebra for Web Information. In Proceedings of the IEEE Advances in Digital Libraries Conference, Santa Barbara, U.S.A., April 1998.
8. Shian-Hua Lin, Chi-Sheng Shih, Meng Chang Chen, et al. Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach. In Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.
9. Sourav S. Bhowmick, W.-K. Ng, E.-P. Lim. Information Coupling in Web Databases. In Proceedings of the 17th International Conference on Conceptual Modelling(ER’98), Singapore, November 16-19, 1998.
10. D. Backman and J. Rubbin, Web log analysis: Finding a Recipe for Success. http://techweb.comp. com/nc/811/811cn2.html, 1997.
11. M.S. Chen, J. Han and P.S. Yu, Data Mining: An Overview from a Database Perspective. IEEE Transaction on Knowledge and Data Engineering, 8:866-833, 1996.
12. J. Han and Y. Fu. Discovery of Multi-level Association Rules. In Proceedings of International Conference on Very Large Databases, pages 420-431, Zurich, Switzerland, Sept. 1995.
13. J. Pitkow, In Search of Reliable Usage Data on the WWW. In Proceedings of the 6th International World Wide Web Conference, Santa Clara, California, April, 1997.
14. World Wide Web Consortium. Document Object Model (DOM) Level 1 Specification. http://www.w3.org/TR/REC-DOM-Level1.
15. K. Wang, H. Liu. Discovering Typical Structures of Documents: A Road Map Approach, ACM SIGR, August 1998.
16. K. Wang, H. Liu, Schema Discovery for Semi structured Data. In Proceedings of International Conference on Knowledge Discovery and Data Mining, Newport Beach, AAAI, Aug. 1997.

AUTHORS

Vishal has obtained his MCA degree in 2006 from Uttar Pradesh Technical University. [U.P] INDIA .His research interest is Information technology. Recently he is doing research from Shri Venkateshwara University Gajraula, Amroha, (U.P.) - 226025, India

Dr. Saurabh Gupta has obtained his Ph.D. Degree from Lucknow University in the year 1999. He has completed his Ph.D. in the area of Computer Engineering. His research interests are, new innovation in e-Governance etc. He has published many of the valuable research papers in various national and international Journals. He is presently working as a State Informatics Officer in National Informatics Centre Shimla (Himachal Pradesh), India.