International Journal of Engineering, Science and Metallurgy

(General article)

WHOWEDA: A New Approach of Web Structure, Content and Usages Mining

Aradhana Mehta[a] , Pawan Bishnoi[a]

[a]CSE Deptt. , OITM Hisar.

Accepted 2 July 2012, Available online 1Sept 2012

Abstract

In this paper, we present an overview of research issues in web mining. The paper discuss about the WHOWEDA and how the warehousing of data is done in the Web. 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 warehousing project called WHOWEDA (Warehouse of Web Data). 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.

Keywords: *Web Data, Ware Housing, WHOWEDA*

1. 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 [1].

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. Thus, the WWW data mining should focus on three issues; web structure mining, web content mining [8] and web usage mining [2,10].

Web structure mining involves mining the web document's structures and links. In, some insight is given on mining structural information on the web. Our initial study [5] has shown that web structure mining is very useful in generating information such visible web documents, luminous web documents and luminous paths; a path common to most of the results returned. In this paper, we have discussed some applications in web data mining and E-commerce where we can use these types of knowledge. Web content mining describes the automatic search of information resources available on-line. Web usage mining includes the data from server access logs, user registration or profiles, user sessions or transactions etc. A survey of some of the emerging tools and techniques for web usage mining is given in [2]. In our discussion here, we focus on the research issues in web data mining with respect to the web warehousing project called WHOWEDA (Warehouse of Web Data).

The key objective of WHOWEDA at the Centre for Advanced Information Systems in Nanyang Technological University, Singapore is to design and implement a web warehouse that materializes and manages useful information from the web to support strategic decision making. We are building a web warehouse [7] using the database approach of managing a web warehouse containing strategic information coupled from the web that may also inter-operate with conventional data warehouses. One of the important areas of our work involves the development of techniques for mining useful information from the web. We would be integrating WHOWEDA with intelligent tools for

information retrieval and extend the data mining techniques to provide a higher level of data organization for unstructured data available on the web.

2. WHOWEDA

In WHOWEDA, we introduced our web data model. It consists of a hierarchy of web objects. The fundamental objects are Nodes and Links, where nodes correspond to HTML text documents and links correspond to hyper-links interconnecting the documents in the WWW. These objects consist of a set of attributes as follows: Nodes = [url, title, format, size, date, text] and link = [source-url, target-url, label, link-type]. In our web warehouse, Web Information Coupling System (WICS) [9] is a database system for managing and manipulating coupled information extracted from the Web. We have defined a set of coupling operators to manipulate the web tables and correlate additional useful and related information [9].
We materialize web data as web tuples stored in web tables. Web tuples, representing directed connecting graphs, are comprised of web objects (Nodes and Links). We associate with each web table a web schema that binds a set of web tuples in a web table. A web schema contains the meta-data that binds a set of web tuples to a web table in the form of connectivities and predicates defined on node and link variables. Connectivities represent structural properties of web tuples by describing possible paths between node variables. Predicates on the other hand specify the additional conditions that must be satisfied by each tuple to be included in the web table. In WICS, a user expresses a web query in the form of a query graph consisting of some nodes and links representing web documents and hyperlinks in those documents, respectively. Each of these nodes and links can have some keywords imposed on them to represent those web documents that contain the given keywords in the documents and/or hyperlinks. When the query graph is posted over the WWW, a set of web tuples each satisfying the query graph are harnessed from the WWW. Thus, the web schema of a table resembles the query graph used to derive the web tuples stored in web table. Note that the results are returned as web tuples. Note that some nodes and links in the query graph may not have keywords imposed. They are called unbound nodes and links, respectively.

3. Web structure mining

Web information retrieval tools make use of only the text on pages, ignoring valuable information contained in links. Web structure mining aims to generate structural summary about web sites and web pages. The focus of structure mining is therefore on link information, which is an important aspect of web data. Given a collection of interconnected web documents, interesting and informative facts describing their connectivity in the web subset can be discovered. We are interested in generating the following structural information from the web tuples stored in the web tables.

Measuring the frequency of the local links in the web tuples in a web table. Local links connect the different web documents residing in the same server. This informs about the web tuples (connected documents) in the web table that have more information about inter-related documents existing at the same server. This also measures the completeness of the web sites in a sense that most of the closely related information are available at the same site. For example, an airline's home page will have more local links connecting the "routing information with air-fares and schedules" than external links.

4. Web bags

We have defined a concept of a web bag in [5] and used web bags for the types of the knowledge discovery discussed above. Informally, a web bag is a web table containing multiple occurrences of identical web tuples. Note that a web tuple is a set of inter-linked documents retrieved from the WWW that satisfies a query graph. A web bag may only be created by projecting some of the nodes from web tuples of a web table using the web project operator. A web project operator is used to isolate the data of interest, allowing subsequent queries to run over a smaller, perhaps more structured web data. Unlike its relational counterpart, a web project operator does not eliminate identical web tuples autonomously. Thus, the projected web table may contain identical web tuples (i.e., a web bag). The duplicate elimination process is initiated explicitly by a user. Autonomous duplicate elimination may hinder the possibility of discovering useful knowledge from a web table. This is due to the fact that such knowledge may only be discovered from web bags.
Using web bags, we discover visible web documents, luminous web documents and luminous paths [5]. Below we define the three types of knowledge. Then we discuss the applications of three types of knowledge, which we are currently working.

Visibility of Web Documents : Visibility of web documents D in a web table W measures the number of different web documents in W that have links to D. We call such documents visible since they are visible in the web table as they are linked by large number of distinct nodes. The significance of a visible node D is that the document D is relatively more important compared to other documents or nodes in W for the given query. In a web table, each node variable may have a set of visible nodes. All of these may not be useful to the user. Thus, we explicitly specify a threshold value to control the search for visible nodes. The visibility threshold indicates that there should exist at least some reasonably substantial evidence of the visibility of instances of the

specified node variable in the web table to warrant the presentation of visible nodes.

Luminosity of Web Documents : Reversing the concepts of visibility , luminosity of a web document D in a web table W measures the number of outgoing links, i..e, the number of other distinct web documents in W that are linked from D. Similar to the determination of visible nodes, we explicitly specify the node variable y based on which luminous nodes are to be discovered and the luminosity threshold.

Applications : One can use luminosity of a web site, displaying a particular or a set of products, to identify the companies that make all those products.

5. Web content mining

Web content mining involves mining web data contents. The open question is what does it mean to mine content from the web? In effect web content mining is the analog of data mining techniques for relational databases since we can expect to find similar types of knowledge from unstructured data residing in web documents. The unstructured nature of web data forces a different approach towards web content mining. The web contains a mix of many different data types such as textual data, image data, audio and video, etc. In WHOWEDA, currently we primarily focus on mining useful information from the web hypertext data. In particular, we consider the following issues of web content mining in the web warehouse context:

Similarity and difference between web content mining in web warehouse context and conventional data mining. In relational database, the data are flat are very well arranged in a tabular structure defined using attributes whose domains are known. In case of web data, documents are totally unstructured and different attributes in documents may have semantically similar meaning across WWW or vice versa.

Selection of type of data in the WWW to do web content mining. Web content mining needs to select useful information before analysis. It is not practical to expect data mining system to search the entire WWW to discover knowledge requested by the user. In our case, we mine based on the meta-data available. Even if scalability argument is ignored, large amount of redundant, uninteresting pieces of information may be returned.

Cleaning of selected data to mine effectively. This is the step after the web data is selected for mining. Before mining, one may need to transform the data into some data model, which is well understood.

Types of knowledge that can be discovered in a web warehouse context. The types of knowledge to be discovered are as follows: generalized relation, characteristic rule, discriminate rule, classification rule, association rule, and deviation rule.

Discovery of types of information hidden in a web warehouse which are useful for decision making. Web data sources being heterogeneous , diverse and unstructured, are difficult to categorize. To perform interactive web content mining. Presentation of discovered knowledge to the users to expedite complex decision making.

Role of WICM (Web Information Coupling Model) aid in web content mining. We have build a coupling system, which brings the results from the web using a query. WICM plays an important role is populating the warehouse as our web data mining is restricted to the results returned in response to a query. Therefore, web coupling plays an important role in selecting of data.

6. Web usage mining

In WHOWEDA, the user initiates a coupling framework to collect related information. For example, a user may be interested in coupling a query graph " to find the hotel information" with the query graph " to find the places of interest" . From this query graph, we can generate some user access pattern of coupling framework. We can generate a rule like " 50% of users who query " hotel" also couple their query with " places of interest" . This information can be used in the warehouse in local coupling; coupling of materialized web tables containing information on hotels with places of interests. Another information that can be of interest is to find coupled concepts from the coupling framework. This can be used in organizing web sites. For example, web documents that provide information on " hotels" should also have hyperlinks to web pages providing information on " places of interest" . These coupled concepts can also be used to design the Warehousing Concept Mart (WCM), discussed in next section.

Conclusion

In this paper, we have discussed some web data mining research issues in context of the web warehousing project called WHOWEDA (Warehouse of Web Data). 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 warehousing project is to design the tools and techniques 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/isf/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.
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.