Research Issues in Web Data Mining

SANJAY MADRIA, SOUROV S BHOWMICK, W. -K NG, E. P. LIM

Center for Advanced Information Systems, School of Applied Science
Nanyang Technological University, Singapore 639798
{askumar, p517026, awkng, aseplim}@ntu.edu.sg

Abstract. In this paper, we present an overview of research issues in web 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 warehousing project called WHOWEDA (*W*are*ho*use of *We*b *Da*ta). We have categorized web data mining into threes 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.

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]. A more recent survey of web query processing has appeared in [23].

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* [2,10,13]. Web structure mining involves mining the web document's structures and links. In [24], 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 (*W*are*ho*use of *We*b *D*at*a).*

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.

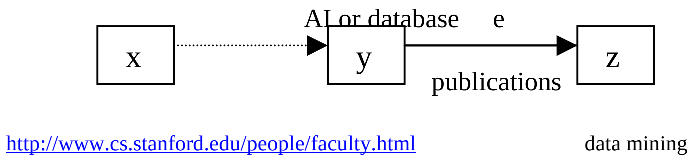
With respect to our web data mining approach, we argue that extracting information from a very small subset of all HTML web pages is also an instance of web data mining. In WHOWEDA, we focus on mining a subset of web pages stored in one or more web tables because we believe that due to the complexity and vastness of the web, mining information from a subset of web stored in the web tables is more feasible option. Our web warehousing approach allows us to do this effectively as we materialize only the results returned in response to a user's query graph.

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].

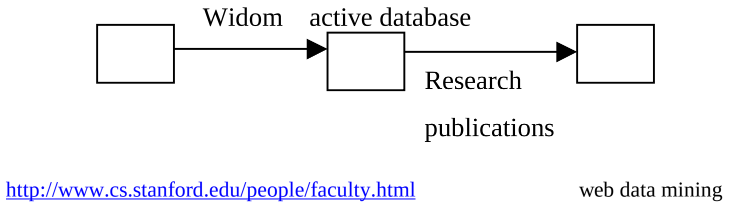
We materialize web data as web tuples stored in web tables. Web tuples, representing directed connecting graphs, are comprised of web objects (Nodes and Links). We associate with each web table a web schema that binds a set of web tuples in a web table. A web schema contains the meta-data that binds 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.

Consider a query to find all data mining related publications by the computer science faculty at Stanford University, starting with the web page http://www.cs.stanford.edu/people/faculty.html. The query above may be expressed as follows :



The above query graph is assigned as schema to the web table generated in response the above query. The schema corresponding to the above query graph can be formally expressed as $<X_n, X_l, C, P>$ where X_n is the set node variables; x,y,z in the example above, X_l is the set of link variables; - (unbound link) and e in the example, C is set of connectivities ; $k_1 \wedge k_2$ where k_1 = x<->y, k_2 = y<e>z and P is a set of predicates as follows : $p_1 \wedge p_2 \wedge p_3 \wedge p_4$ such that p_1 (x) = [x.url EQUALS http://www.cs.standford.edu/people/faculty.html], p_2 (e) = [e.label CONTAINS "publications"], p_3 (y) = [y.text contains "AI or database"], p_4(z) = [z.text CONTAINS "data mining"].

The query returns all web tuples satisfying the web schema given above. These web tuples contain the faculty page, the faculty member's page that should contain the word such "AI or database" and the respective publications page if it contains the word "data mining". Thus, many instances of the query graph shown above will be returned as web tuples. We show one of the instance of the above query graph below.



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.

Note that our web warehouse is populated using a query graph initiated by the user. If the results returned (i..e, web tuples) are having more local links then we can know that the query basically crawl some particular web sites locally. In such a case, for next execution of such a query, one can optimize the query graph to start crawling with those sites directly. This case, for example, may arise when the web crawler passes through many unbound nodes and links, but actual web pages are then found at single site. Note that in many cases the information requested may be found at many different servers across the network. Identifying local links depicts the fact that integrated information is also available at some particular web site, but may be in different files. Thus, such an information can reduce the query execution time over the internet.

- Measuring the frequency of web tuples in a web table containing links which are interior; links which are within the same document. This measures a web document's ability to cross-reference other related web pages within the same document. This also measures the flow of the web documents. For example, a news-paper should always refer to other news items locally (within the same news-paper). This information depicts that the relevant information is available within the same file.
- Measuring the frequency of web tuples in a web table that contains links that are global; links which span different web sites. This measures the visibility of the web documents and ability to relate similar or related documents across different sites. For example, research documents related to "semi-structured data" will be available at many sites and such sites should be visible to other related sites by providing cross references by the popular phrases such as "more related links". Also, in case of a document like a research paper, it should have more external links as it should refer to other related papers. This expresses a research paper's ability to cross-reference other related work.
- Measuring the frequency of identical web tuples that appear in a web table or among the web tables. This measures the replication of web documents across the web warehouse and may help in identifying, for example, the mirrored sites. This information concludes that some web pages provide integrated information on various topics. We have also used duplicate web tuples to identify, for example, visible web pages etc. We discuss this issue in next subsection.

- On average, we may need to find how many web tuples are returned in response to a query on some popular phrases such as “Bio-science” with respect to queries containing keywords like “earth-science”. This can give an estimation of the results returned in response to some popular queries. This also gives an indication whether we should store the results of such a query in our warehouse for further reference.
- Another interesting issue is to discover the nature of the hierarchy or network of hyperlinks in the web sites of a particular domain. For example, with respect URLs with domains like .edu, one would like to know how most of the web sites are designed with respect to information flow in educational institutes. What is the flow of the information they provide and how are they related conceptually. Is it possible to extract a conceptual hierarchical information for designing web sites of a particular domain. This may help in generalizing the flow of information in web sites representing information in some particular domain. This will help for example in building a common web schema or wrappers for educational institutes. Thus it can make query processing easier.
- What is the in-degree and out-degree of each node (web document)? What is the meaning of high and low in- and out-degrees? For example, a high in-degree may be a sign of a very popular web site or document. Similarly, a high out-degree may be a sign of luminous web site. Out-degree also measures a site's connectivity. We discuss these issues in next sub-section.
- If a web page is directly linked to another web page or are near to each other then we would like to discover the relationships among those web pages. These relationships might be of the following types. The two web pages might be related by synonyms or ontology or having similar topics, both the web pages are in the same server and in that case both the pages may be authored by the same person.

While the above information is discovered at the inter-document level, web structure mining can also have another direction - discovering the structure of web documents themselves. Web document structure mining can be used to reveal the structure (schema) of web pages. While this would be useful for navigational purpose and several other operations such comparing and integrating web page schemes can be made possible. This type of structure mining would facilitate web document classification and clustering on the basis of structure. It will also contribute towards introducing database techniques for accessing information in web pages by providing a reference schema. The availability of semantic markup language XML [14] has made it possible to identify tree structure within such documents. These structures can be compared using n-ary tree matching algorithms. Schema integration can be carried out by generating representative structures such as centroid and spanning trees. Related work on schema discovery of semi-structured documents includes [15,16] which uses the Object Exchange Model (OEM) devised in LORE [18] and is similar to approach of using representative objects in [19]. Another work [17] uses OEM but derives a type hierarchy using measures similar to support and confidence encountered earlier, to represent the inherent structure of large collections of semi-structured data.

3.1 Web Bags

Most of the search engines fail to handle the following knowledge discovery goals:

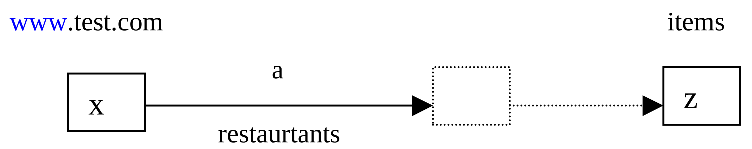
- From the query's result returned by search engines, a user may wish to locate the most visible web sites [6] or documents for reference. That is, many paths (high fan in) can reach that sites or documents. Presently, he may only do so manually by visiting the documents in the query result and then manually follow each links in the web documents and then download the visible documents as files on user's hard disk for future reference. Nevertheless, this method is tedious.
- Reversing the concept of visibility, a user may wish to locate the most luminous web sites [6] or documents for reference. That is, web sites or documents which have the most number of outgoing links. Currently, he may locate this information by manually visiting each web documents.
- Furthermore, a user may wish to find out the most traversed path for a particular query result. This is important since it helps the user to identify the set of most popular interlinked web documents that have been traversed frequently to obtain the query result. Presently, he may only do so by visiting each document in the search result and compare their link information. This method is time consuming.

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.

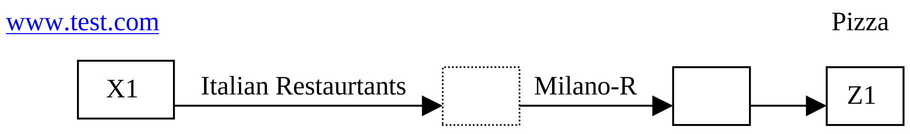
Using web bags, we discover visible web documents, luminous web documents and luminous paths [5]. Below we define the three types of knowledge. Then we discuss the applications of three types of knowledge, which we are currently working.

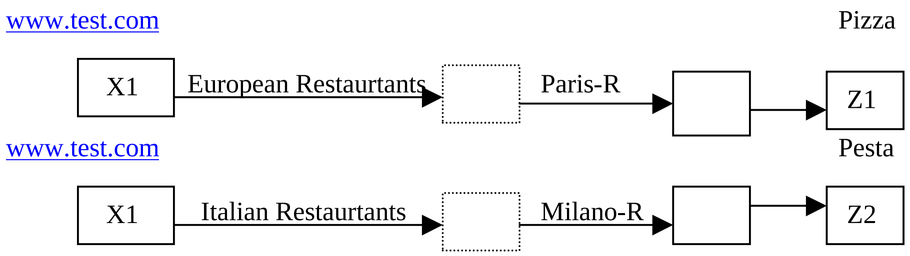
Visibility of Web Documents : Visibility of web documents D in a web table W measures the number of different web documents in W that have links to D. We call such documents visible since they are visible in the web table as they are linked by large number of distinct nodes. The significance of a visible 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.

Applications : Consider a query graph involving some keywords such as " types of restaurants" and "items" given below, where dotted lines implies unbound node and link. We assume that such a site is there on WWW which provides a list of types of restaurants (i.e., Italian, Asian, etc.) which further have names of those restaurants. We also assume that there is a web site which provides list of items for all types of restaurants.



The results returned in response to the query graph imposing such predicates in our web warehouse system will return the instances of restaurants selling different items. For example, the three web tuples corresponding to the query graph are as given below.



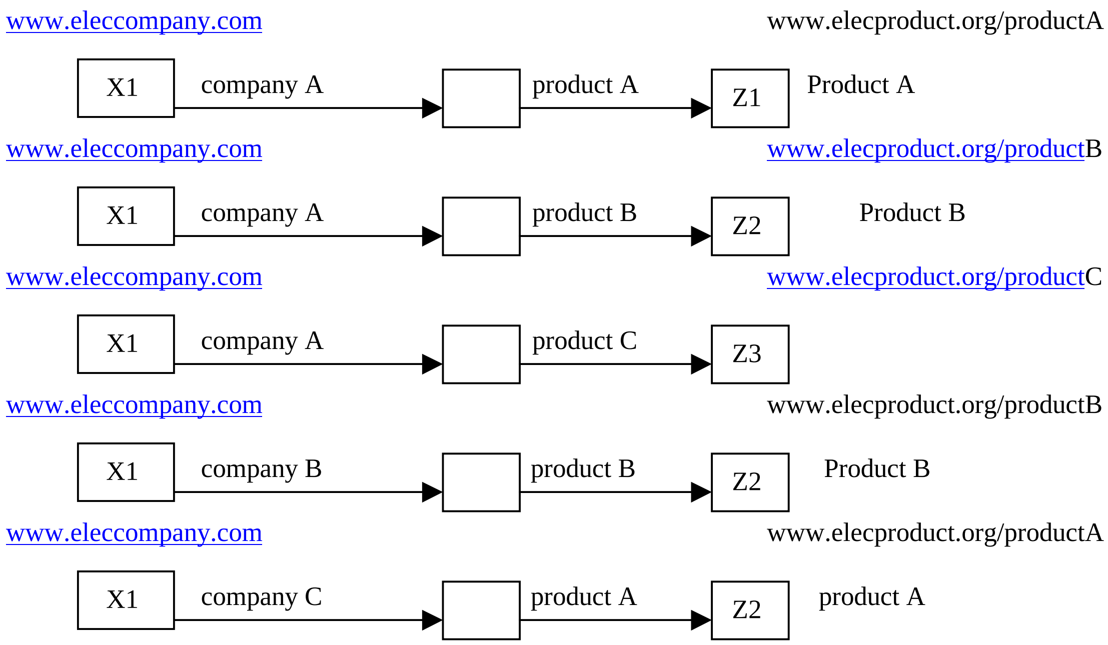


From the results returned, we can find the most visible web pages by providing very high visibility threshold [see [5] for further reference). Assume that this gives Z1 as the most visible web page (having more incoming links from different URLs) which has details about pizza. This can give an estimate about the different restaurants which sell pizzas. By lowering the visibility-threshold, we can get another set of visible web pages, and assume that this time we get the set as {Z1, Z2} where Z2 is an instance of a web-page which provides details of Pasta. Note that it is possible that some restaurants can sell both pizza and pasta. By comparing the set of different URLs corresponding to the restaurants, we can derive the association rules such "out of 80% of restaurants which offer pizza to their customers, 40% also provide pasta. Further, we can cluster (group) these restaurants according to type and can generate rules like out of 80% of restaurants which sell pizza, 40% which sell pasta also are of Italian types.

Consider another example where a new business venture wants to do some analysis of their web sites which display products for buying. By finding the visibility of its web site with respect to other web sites selling such (or related) products, the company can find ways to redesign (including changes in product's price etc.) its web site to improve visibility. For example, if a web site sells PC monitors, they must be providing links to web sites which sell CPU. Thus, if a web site finds that its visibility is lower in comparison to other web sites selling CPUs then the web site needs to improve in terms of design, products, etc.

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. This will given an estimate of the type that a company whenever it makes a product "A" also makes a set of products "B and C". Note that a company can make a product B and/or C only with out necessarily making a product A or it may be possible that a certain percentage of companies demonstrate such rules. We want to generate association rules such as X% of all the electric companies which makes a product "A", Y% of them also makes a set of other products "B and C" (support). Also, we can generate a rule like whenever a company makes a product, it also makes certain other products, for example, X% of companies which make a product A may also make a product B and C (confidence). Such rules help a new electric company in taking a decision such as the set of products the company should start manufacturing together.

Consider the following web tuples in a web table.



Note that in above example, certain companies (20%) if they make a product A also make products B and C. However, the company C makes only the product A. That is, 40% of companies which make a product A , 20% of them also make products B and C.

Luminous Paths: A web query result in WHOWEDA is set of inter-linked web tuples materialized in the form of web table. Luminous paths in a web table is set of inter-linked nodes (paths) which occurs some number of times across tuples in the web table. That is, occurrences of this set of inter-linked nodes is high compared to the total number of web tuples in the web table. An implication is that in order to couple the query results from the WWW, most of the web tuples in the web table has to traverse the luminous paths.

Applications : Luminous paths can be used to optimize the visualization of query results. Once the results are returned, one needs to browse the nodes (web pages) in the set of luminous paths only once. For example, it may be possible that between two web pages there may exists two paths such that one is a subset of another. In that case, common paths (web pages) need to browse only once.

Another interesting application is to find whether two given queries are similar. Consider that two web tables T_1 and T_2 corresponding to two query graphs Q_1 and Q_2. If we find that sets of luminious paths in the two web tables have common sets of luminous paths or sub-paths then we can infer that the corresponding query graphs are similar. We would also like to find the similar relationships; that is, whether, they are conceptually related or the keywords present in two web pages are synonyms to each other, or they are topically related.

4 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. For example, one web site could display the price of same car in numeric figure others may do in words. An attribute may have an atomic value in one document, but a set of values in other documents. In order to do content mining, one must first resolve the problems of semantic integration across web documents.

- 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. The user must therefore be provided the facility to identify a subset of the web, which pertains to the domain of the knowledge discovery task. Then, depending on the specific kind of knowledge to be mined another level of data selection must be carried out to extract relevant data into a suitable representative model.
- 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. For example, in our case, we transform the web data in the form of web tuples consisting of nodes and links and having keywords specified over them. We also use web algebraic operators to filter out the irrelevant information.
- 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 [11]. Do these data mining techniques applicable to web data mining and if yes, how? For example, we are interested in generating the following types of rules: 80% of web tuples (i..e, web pages) in response to a "travel information query from Hong Kong to Macau" suggest that popular means of traveling is by ferry.
- 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. In many cases, the user would be even more unsure about the knowledge hidden beneath the contents of a document than that in a database. An interactive and iterative process is therefore necessary to enable exploratory data mining. A suitable data mining query language is one of the means to materliaze such a user-mediated process. In our WHOWEDA project, we are building a query language which support mining.
- To perform interactive web content mining. Presentation of discovered knowledge to the users to expedite complex decision making. For example, a query interface is often necessary to specify the interesting set of data to be studied, the kind of rules to be discovered, etc. A graphical user interface is

helpful for interactive mining of multiple-level rules [12] because it facilitates interactive modification of the threshold values, warehouse concept mart (discussed later), concept levels, output styles and formats.

- Role of WICM (web information coupling model [9]) 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.

5 Web Usage Mining

Web usage mining is the discovery of user access patterns from web server logs, which maintain an account of each user browsing activities. Web servers automatically generate large data stored in sever referred as logs containing information about the user profile, access pattern for pages, etc. This can provide information that can be used for efficient and effective web site management and the user behavior. Apart from finding paths traversed frequently by users as a series of URLs, associations indicate which sites are likely to be visited together can also be derived. One particular approach mentioned in [11] using "maximal forward references" obtained by filtering out "backward references" from traversal subsequences in log data to extract frequently occurring consecutive subsequences. This leads to "maximal reference sequence" which are those frequent subsequences that are not subset of others. The problem is similar to order version of finding large itemsets in transaction databases [20,21] discussed under mining of association rules. An improvement over maximal forward references which considers backward references is discussed in [22] and uses a transaction model for data extracted from server access logs to discover sequential patterns. This approach combines all the entries for a user in a server log into a single transaction using clustering. Association rules and sequential patterns can then be mined from the grouped transactions.

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.

6 Warehouse Concept Mart

Knowledge discovery in web data becomes more and more complex due to the large number of data on WWW. We are building the concept hierarchies involving web data to use them in knowledge discovery. We call such collection of concept hierarchies a Warehouse Concept Mart (WCM). The concept mart is build by extracting and generalizing terms from web documents to represent classification knowledge of a given class hierarchy. For unclassified words, they can be clustered based on their common properties. Once the clusters are decided, the keywords can be labeled with their corresponding clusters, and common features of the terms are summarized to form the concept description. We can associate a weight at each level of concept marts to evaluate the importance of a term with respect to the concept level in the concept hierarchy. The concept marts can be used for the following:

1. Intelligent answering of web queries

Knowledge discovery using the warehouse concept mart facilitates querying web data and intelligent query answering in web warehousing system WHOWEDA. A user can supply the threshold for a given key word in the warehouse concept mart and the words with the threshold above the given value can be taken into consideration when answering the query. The query can also be answered using different levels of concept in the warehouse concept mart [3] or can provide approximate answers [4].

Another interesting idea is to provide the user some knowledge in framing the global coupling query graph. For example, If a user frame a query graph to find some information about “Database system”, he can be supplied some related concepts like “ORACLE” in case he would like to pose a query by coupling “Database system” with “ORACLE”.

2. Web Data Mining and Concept Mart

Warehouse Concept Mart (WCM) can be used for web data or content mining. In web content mining, we make use of the warehouse concept mart in generating some of the useful knowledge. We are mining association rules techniques to mine the association between words appearing in the concept mart at various levels and in the web tuples returned as the result of a query. Mining knowledge at multiple levels may help WWW users to find some interesting rules that are difficult to be discovered otherwise. A knowledge discovery process may climb up and step down to different concepts in the warehouse concept mart's level with user's interactions and instructions including different threshold values.

Another application of the warehouse concept mart is in web structure mining. For example, as mentioned before, we can capture the flow of web sites of particular domain and based on that, we may be able to generalize the structure of web sites of particular domain.

7 Conclusions

In this paper, we have discussed some web data mining research issues in context of the web warehousing project called WHOWEDA (*Ware*house of *We*b *Da*ta) at Center for Advanced Information Systems, Nanyang Technological University, Singapore. 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/is-f/library-f/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.

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 Semistructured Data. In Proceedings of International Conference on Knowledge Discovery and Data Mining, Newport Beach, AAAI, Aug. 1997.

17. S. Nestorov, S. Abiteboul, R. Motwani. Inferring Structure in Semistructured Data. In Proceedings of International Workshop on Management of Semistructured Data, 1997.

18. J. Mchugh, S. Abiteboul, R. Goldman, D. Quass, J. Widom, Lore : A Database Management System for Semistructured Data, SIGMOD Record, 26(3):54-66, September, 1997.
19. S. Nestorov , J. Ullman, J. Widom, S. Chawathe, Representative Objects: Concise Representations of Semistructured, Hierarchical Data. In proceedings of IEEE International Conference on Data Engineering, pp. 79-90, Birmingham, U.K., 1997.
20. H. Mannila, Methods and Problems in Data Mining. In Proceedings of International Conference on Database theory, Delphi, Greece, January 1997.
21. R. Aggarwal, R. Srikant, Fast Algorithms for Mining Association Rules. In proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.
22. B. Mobasher, N. Jain, J. Han, J. Srivastava, Web Mining : Pattern Discovery From World Wide Web Transactions. In Proceedings of the 9th IEEE International Conference on Tools with AI (ICTAI,97), Nov. 1997.
23. D. Florescu, A. Levy, A. Mendelzon, Database Techniques for the World Wide Web, A Survey, SIGMOD Record, 1998.
24. Ellen, Spertus, ParaSite : Mining Structural Information on the Web. In proceedings of 6th International WWW Conference , April, 1997.