Web Mining: Information and Pattern Discovery on the World Wide Web *

R. Cooley, B. Mobasher, and J. Srivastava
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455, USA

Abstract

Application of data mining techniques to the World Wide Web, referred to as Web mining, has been the focus of several recent research projects and papers. However, there is no established vocabulary, leading to confusion when comparing research efforts. The term Web mining has been used in two distinct ways. The first, called Web content mining in this paper, is the process of information discovery from sources across the World Wide Web. The second, called Web usage mining, is the process of mining for user browsing and access patterns. In this paper we define Web mining and present an overview of the various research issues, techniques, and development efforts. We briefly describe WEBMINER, a system for Web usage mining, and conclude this paper by listing research issues.

1 Introduction

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in find the desired information resources, and to track and analyze their usage patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge. *Web mining* can be broadly defined as the discovery and analysis of useful information from the World Wide Web. This describes the automatic search of information resources available on-line, i.e. *Web content mining*, and the discovery of user access patterns from Web servers, i.e., *Web usage mining.*

In this paper, we provide an overview of tools, techniques, and problems associated with both dimensions. We present a taxonomy of Web mining, and place various aspects of Web mining in their proper context. There are several important issues, unique to the Web paradigm, that come into play if sophisticated types of analyses are to be done on server side data collections. These include integrating various data sources such as server access logs, referrer logs, user registration or profile information; resolving difficulties in the identification of users due to missing unique key attributes in collected data; and the importance of identifying user sessions or transactions from usage data, site topologies, and models of user behavior. We devote the main part of this paper to the discussion of issues and problems that characterize Web usage mining. Furthermore, we survey some of the emerging tools and techniques, and identify several future research directions.

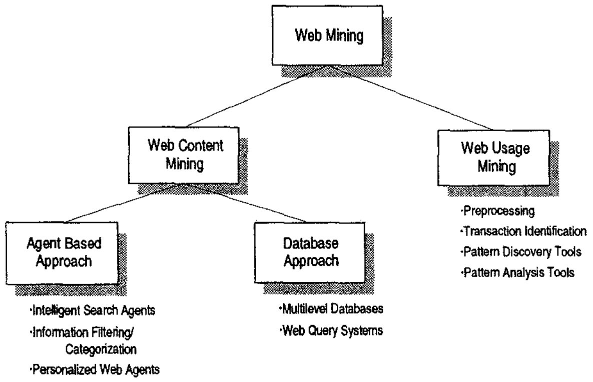

Figure 1: Taxonomy of Web Mining

2 A Taxonomy of Web Mining

In this section we present a taxonomy of Web mining, i.e. Web content mining and Web usage mining. We also describe and categorize some of the recent work and the related tools or techniques in each area. This taxonomy is depicted in Figure 1.

This work was supported in part by NSF grant ASC-9554517

1082-3409/97 $10.00 © 1997 IEEE

2.1 Web Content Mining

The lack of structure that permeates the information sources on the World Wide Web makes automated discovery of Web-based information difficult. Traditional search engines such as Lycos, Alta Vista, WebCrawler, ALIWEB [29], MetaCrawler, and others provide some comfort to users, but do not generally provide structural information nor categorize, filter, or interpret documents. A recent study provides a comprehensive and statistically thorough comparative evaluation of the most popular search engines [32].

In recent years these factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent Web agents, and to extend data mining techniques to provide a higher level of organization for semi-structured data available on the Web. We summarize some of these efforts below.

2.1.1 Agent-Based Approach. Generally, agent-based Web mining systems can be placed into the following three categories:

Intelligent Search Agents: Several intelligent Web agents have been developed that search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information. Agents such as Harvest [6], FAQ-Finder [19], Information Manifold [27], OCCAM [30], and ParaSite [51] rely either on pre-specified domain information about particular types of documents, or on hard coded models of the information sources to retrieve and interpret documents. Agents such as ShopBot [14] and ILA (Internet Learning Agent) [42] interact with and learn the structure of unfamiliar information sources. ShopBot retrieves product information from a variety of vendor sites using only general information about the product domain. ILA learns models of various information sources and translates these into its own concept hierarchy.

Information Filtering/Categorization: A number of Web agents use various information retrieval techniques [17] and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them [5, 9, 34, 55, 53]. HyPursuit [53] uses semantic information embedded in link structures and document content to create cluster hierarchies of hypertext documents, and structure an information space. BO (Bookmark Organizer) [34] combines hierarchical clustering techniques and user interaction to organize a collection of Web documents based on conceptual information.

Personalized Web Agents: This category of Web agents learn user preferences and discover Web information sources based on these preferences, and those of other individuals with similar interests (using collaborative filtering). A few recent examples of such agents include the WebWatcher [3], PAINT [39], Syskill & Webert [41], GroupLens [47], Firefly [49], and others [4]. For example, Syskill & Webert utilizes a user profile and learns to rate Web pages of interest using a Bayesian classifier.

2.1.2 Database Approach. Database approaches to Web mining have focused on techniques for organizing the semi-structured data on the Web into more structured collections of resources, and using standard database querying mechanisms and data mining techniques to analyze it.

Multilevel Databases: The main idea behind this approach is that the lowest level of the database contains semi-structured information stored in various Web repositories, such as hypertext documents. At the higher level(s) meta data or generalizations are extracted from lower levels and organized in structured collections, i.e. relational or object-oriented databases. For example, Han, et. al. [56] use a multi-layered database where each layer is obtained via generalization and transformation operations performed on the lower layers. Kholsa, et. al. [25] propose the creation and maintenance of meta-databases at each information providing domain and the use of a global schema for the meta-database. King & Novak [26] propose the incremental integration of a portion of the schema from each information source, rather than relying on a global heterogeneous database schema. The ARANEUS system [40] extracts relevant information from hypertext documents and integrates these into higher-level ***derived Web Hypertexts*** which are generalizations of the notion of database views.

Web Query Systems: Many Web-based query systems and languages utilize standard database query languages such as SQL, structural information about Web documents, and even natural language processing for the queries that are used in World Wide Web searches. W3QL [28] combines structure queries, based on the organization of hypertext documents, and content queries, based on information retrieval techniques. WebLog [31] Logic-based query language for restructuring extracts information from Web information sources. Lorel [46] and UnQL [7, 8] query heterogeneous and semi-structured information on the Web using a labeled graph data model. TSIMMIS [10] extracts data from heterogeneous and semi-structured information sources and correlates them to generate an integrated database representation of the extracted information.

2.2 Web Usage Mining

Web usage mining is the automatic discovery of user access patterns from Web servers. Organizations collect large volumes of data in their daily operations, generated automatically by Web servers and collected in server access logs. Other sources of user information include *referrer logs* which contain information about the referring pages for each page reference, and user registration or survey data gathered via CGI scripts.

Analyzing such data can help organizations determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among other things. It can also provide information on how to restructure a Web site to create a more effective organizational presence, and shed light on more effective management of workgroup communication and organizational infrastructure. For selling advertisements on the World Wide Web, analyzing user access patterns helps in targeting ads to specific groups of users.

Most existing Web analysis tools [16, 23, 37] provide mechanisms for reporting user activity in the servers and various forms of data filtering. Using such tools it is possible to determine the number of accesses to the server and to individual files, the times of visits, and the domain names and URLs of users. However, these tools are designed to handle low to moderate traffic servers, and usually provide little or no analysis of data relationships among the accessed files and directories within the Web space.

More sophisticated systems and techniques for discovery and analysis of patterns are now emerging. These tools can be placed into two main categories, as discussed below.

2.2.1 Pattern Discovery Tools. The emerging tools for user pattern discovery use sophisticated techniques from AI, data mining, psychology, and information theory, to mine for knowledge from collected data. For example, the WEBMINER system [12, 36] introduces a general architecture for Web usage mining. WEBMINER automatically discovers association rules and sequential patterns from server access logs. In [11] algorithms are introduced for finding *maximal forward references* and *large reference sequences*. These can, in turn be used to perform various types of user traversal path analysis, such as identifying the most traversed paths thorough a Web locality. Pirolli et. al. [43] use information foraging theory to combine path traversal patterns, Web page typing, and site topology information to categorize pages for easier access by users.

2.2.2 Pattern Analysis Tools. Once access patterns have been discovered, analysts need the appropriate tools and techniques to understand, visualize, and interpret these patterns, e.g. the WebViz system [45]. Others have proposed using OLAP techniques such as data cubes for the purpose of simplifying the analysis of usage statistics from server access logs [15]. The WEBMINER system [36] proposes an SQL-like query mechanism for querying the discovered knowledge (in the form of association rules and sequential patterns). These techniques and others are further discussed in the subsequent sections.

3 Pattern Discovery from Web Transactions

As discussed in section 2.2, analysis of how users are accessing a site is critical for determining effective marketing strategies and optimizing the logical structure of the Web site. Because of many unique characteristics of the client-server model in the World Wide Web, including differences between the physical topology of Web repositories and user access paths, and the difficulty in identification of unique users as well as user sessions or transactions, it is necessary to develop a new framework to enable the mining process. Specifically, there are a number of issues in preprocessing data for mining that must be addressed before the mining algorithms can be run. These include developing a model of access log data, developing techniques to clean/filter the raw data to eliminate outliers and/or irrelevant items, grouping individual page accesses into semantic units (i.e. transactions), integration of various data sources such as user registration information, and specializing generic data mining algorithms to take advantage of the specific nature of access log data.

3.1 Preprocessing Tasks

The first preprocessing task is *data cleaning.* Techniques to clean a server log to eliminate irrelevant items are of importance for any type of Web log analysis, not just data mining. The discovered associations or reported statistics are only useful if the data represented in the server log gives an accurate picture of the user accesses of the Web site. Elimination of irrelevant items can be reasonably accomplished by checking the suffix of the URL name. For instance, all log entries with filename suffixes such as, gif, jpeg, GIF, JPEG, jpg, JPG, and map can be removed.

A related but much harder problem is determining if there are important accesses that are not recorded in the access log. Mechanisms such as local caches and proxy servers can severely distort the overall picture of user traversals through a Web site. Current methods to try to overcome this problem include the use of cookies, cache busting, and explicit user registration. As detailed in [44], none of these methods are without serious drawbacks. Cookies can be deleted by the user, cache busting defeats the speed advantage that caching was created to provide and can be disabled, and user registration is voluntary and users often provide false information. Methods for dealing with the caching problem include using site topology or referrer logs, along with temporal information to infer missing references.

Another problem associated with proxy servers is that of user identification. Use of a machine name to uniquely identify users can result in several users being erroneously grouped together as one user. An algorithm presented in [43] checks to see if each incoming request is reachable from the pages already visited. If a page is requested that is not directly linked to the previous pages, multiple users are assumed to exist on the same machine. In [12], user session lengths determined automatically based on navigation patterns are used to identify users. Other heuristics involve using a combination of IP address, machine name, browser agent, and temporal information to identify users [44].

The second major preprocessing task is *transaction identification.* Before any mining is done on Web usage data, sequences of page references must be grouped into logical units representing Web transactions or user sessions. A user session is all of the page references made by a user during a single visit to a site. Identifying user sessions is similar to the problem of identifying individual users, as discussed above. A transaction differs from a user session in that the size of a transaction can range from a single page reference to all of the page references in a user session, depending on the criteria used to identify transactions. Unlike traditional domains for data mining, such as point of sale databases, there is no convenient method of clustering page references into transactions smaller than an entire user session. This problem has been addressed in [11] and [12].

3.2 Discovery Techniques on Web Transactions

Once user transactions or sessions have been identified, there are several kinds of access pattern mining that can be performed depending on the needs of the analyst, such as path analysis, discovery of association rules and sequential patterns, and clustering and classification.

There are many different types of graphs that can be formed for performing *path analysis*, since a graph represents some relation defined on Web pages (or other objects). The most obvious is a graph representing the physical layout of a Web site, with Web pages as nodes and hypertext links between pages as directed edges. Other graphs could be formed based on the types of Web pages with edges representing similarity between pages, or creating edges that give the number of users that go from one page to another [43]. Most of the work to date involves determining frequent traversal patterns or large reference sequences from the physical layout type of graph. Path analysis could be used to determine most frequently visited paths in a Web site. Other examples of information that can be discovered through path analysis are:

- 70% of clients who accessed `/company/product2` did so by starting at `/company` and proceeding through `/company/new`, `/company/products`, and `/company/product1`;
- 80% of clients who accessed the site started from `/company/products`; or
- 65% of clients left the site after four or less page references.

The first rule suggests that there is useful information in `/company/product2`, but since users tend to take a circuitous route to the page, it is not clearly marked. The second rule simply states that the majority of users are accessing the site through a page other than the main page (assumed to be `/company` in this example) and it might be a good idea to include directory type information on this page if it is not there already. The last rule indicates an attrition rate for the site. Since many users don't browse further than four pages into the site, it would be prudent to ensure that important information is contained within four pages of the common site entry points.

Association rule discovery techniques [1, 48] are generally applied to databases of transactions where each transaction consists of a set of items. In such a framework the problem is to discover all associations and correlations among data items where the presence of one set of items in a transaction implies (with a certain degree of confidence) the presence of other items. In the context of Web usage mining, this problem amounts to discovering the correlations among references to various files available on the server by a given client. Each transaction is comprised of a set of URLs accessed by a client in one visit to the server. For example, using association rule discovery techniques we can find correlations such as the following:

- 40% of clients who accessed the Web page with URL `/company/product1`, also accessed `/company/product2`; or
- 30% of clients who accessed `/company/special`, placed an online order in `/company/product1`.

Since usually such transaction databases contain extremely large amounts of data, current association rule discovery techniques try to prune the search space according to *support* for items under consideration. Support is a measure based on the number of occurrences of user transactions within transaction logs.

Discovery of such rules for organizations engaged in electronic commerce can help in the development of effective marketing strategies. But, in addition, association rules discovered from WWW access logs can give an indication of how to best organize the organization's Web space.

The problem of discovering *sequential patterns* [35, 52] is to find inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. In Web server transaction logs, a visit by a client is recorded over a period of time. The time stamp associated with a transaction in this case will be a time interval which is determined and attached to the transaction during the data cleaning or transaction identification processes. The discovery of sequential patterns in Web server access logs allows Web-based organizations to predict user visit patterns and helps in targeting advertising aimed at groups of users based on these patterns. By analyzing this information, the Web mining system can determine temporal relationships among data items such as the following:

- 30% of clients who visited `/company/products`, had done a search in Yahoo, within the past week on keyword w; or
- 60% of clients who placed an online order in `/company/product1`, also placed an online order in `/company/product4` within 15 days.

Another important kind of data dependency that can be discovered, using the temporal characteristics of the data, are similar time sequences. For example, we may be interested in finding common characteristics of all clients that visited a particular file within the time period $[t_1, t_2]$. Or, conversely, we may be interested in a time interval (within a day, or within a week, etc.) in which a particular file is most accessed.

Discovering *classification rules* [20, 54] allows one to develop a profile of items belonging to a particular group according to their common attributes. This profile can then be used to classify new data items that are added to the database. In Web usage mining, classification techniques allow one to develop a profile for clients who access particular server files based on demographic information available on those clients, or based on their access patterns. For example, classification on WWW access logs may lead to the discovery of relationships such as the following:

- clients from state or government agencies who visit the site tend to be interested in the page `/company/product1`; or
- 50% of clients who placed an online order in `/company/product2`, were in the 20-25 age group and lived on the West Coast.

Clustering analysis [24, 38] allows one to group together clients or data items that have similar characteristics. Clustering of client information or data items on Web transaction logs, can facilitate the development and execution of future marketing strategies, both online and off-line, such as automated return mail to clients falling within a certain cluster, or dynamically changing a particular site for a client, on a return visit, based on past classification of that client.

4 Analysis of Discovered Patterns

The discovery of Web usage patterns, carried out by techniques described earlier, would not be very useful unless there were mechanisms and tools to help an analyst better understand them. Hence, in addition to developing techniques for mining usage patterns from Web logs, there is a need to develop techniques and tools for enabling the analysis of discovered patterns. These techniques are expected to draw from a number of fields including statistics, graphics and visualization, usability analysis, and database querying. In this section we provide a survey of the existing tools and techniques. Usage analysis of Web access behavior being a very new area, there is very little work in it, and correspondingly this survey is not very extensive.

Visualization has been used very successfully in helping people understand various kinds of phenomena, both real and abstract. Hence it is a natural choice for understanding the behavior of Web users. Pitkow, et al [45] have developed the WebViz system for visualizing WWW access patterns. A *Web path paradigm* is proposed in which sets of server log entries are used to extract subsequences of Web traversal patterns called *Web paths.* WebViz allows the analyst to selectively analyze the portion of the Web that is of interest by filtering out the irrelevant portions. The Web is visualized as a directed graph with cycles, where nodes are pages and edges are (inter-page) hyperlinks.

On-Line Analytical Processing (OLAP) is emerging as a powerful paradigm for strategic analysis of databases in business settings. It has been recently demonstrated that the functional and performance needs of OLAP require that new information structures be designed. This has led to the development of the *data cube* information model [18], and techniques for its efficient implementation [2, 22, 50]. Recent work [15] has shown that the analysis needs of Web usage data have much in common with those of a data warehouse, and hence OLAP techniques are quite applicable. The access information in server logs is modeled as an append-only history, which grows over time. Since the size of server logs grows quite rapidly, it may not be possible to provide on-line analysis of all of it. Therefore, there is a need to summarize the log data, perhaps in various ways, to make its on-line analysis feasible. Making portions of the log selectively (in)visible to various analysts may be required for security reasons.

One of the reasons attributed to the great success of relational database technology has been the existence of a high-level, declarative, query language, which allows an application to express *what conditions must be satisfied by the data it needs*, rather than having to specify *how to get the required data.* Given the large number of patterns that may be mined, there appears to be a definite need for a mechanism to specify the focus of the analysis. Such focus may be provided in at least two ways. First, constraints may be placed on the database (perhaps in a declarative language) to restrict the portion of the database to be mined for, e.g. [36]. Second, querying may be performed on the knowledge that has been extracted by the mining process, in which case a language for querying knowledge rather than data is needed. An SQL-like querying mechanism has been proposed for the WEBMINER system [36]. For example, The query

```
SELECT association-rules(A*B*C*)
FROM   log.data
WHERE  date >= 970101 AND domain = "edu"
  AND support = 1.0 AND confidence = 90.0
```

extracts the rules involving the ".edu" domain after Jan 1, 1997, which start with URL A, and contain B and C in that order, and that have a minimum support of 1 % and a minimum confidence of 90 %.

5 Web Usage Mining Architecture

We have developed a general architecture for Web usage mining which is presented in [13] and [36]. The WEBMINER is a system that implements parts of this general architecture. The architecture divides the Web usage mining process into two main parts. The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes preprocessing, transaction identification, and data integration components. The second part includes the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine. The overall architecture for the Web mining process is depicted in Figure 2.

Data cleaning is the first step performed in the Web usage mining process. Some low level data integration tasks may also be performed at this stage, such as combining multiple logs, incorporating referrer logs, etc. After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. The goal of transaction identification is to create meaningful clusters of references for each user. The task of identifying transactions is one of either *dividing* a large transaction into multiple smaller ones or *merging* small transactions into fewer larger ones. The input and output transaction formats match so that any number of modules to be combined in any order, as the data analyst sees fit.

Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data mining task. For instance, the format of the data for the association rule discovery task may be different than the format necessary for mining sequential patterns. Finally, a query mech-

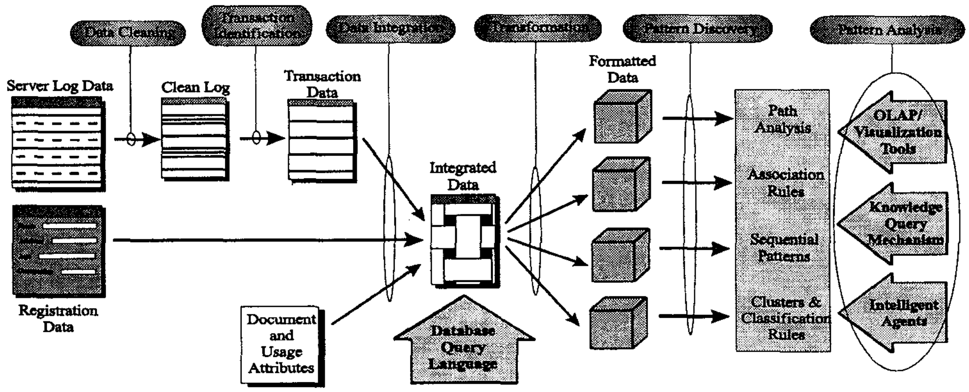

Figure 2: A General Architecture for Web Usage Mining

anism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints. For more details on the WEBMINER system refer to [13, 36].

6 Research Directions

The techniques being applied to Web content mining draw heavily from the work on information retrieval, databases, intelligent agents, etc. Since most of these techniques are well known and reported elsewhere, we have focused on Web usage mining in this survey instead of Web content mining. In the following we provide some directions for future research.

6.1 Data Pre-Processing for Mining

Web usage data is collected in various ways, each mechanism collecting attributes relevant for its purpose. There is a need to pre-process the data to make it easier to mine for knowledge. Specifically, we believe that issues such as instrumentation and data collection, data integration and transaction identification need to be addressed.

Clearly improved data quality can improve the quality of any analysis on it. A problem in the Web domain is the inherent conflict between the analysis needs of the analysts (who want more detailed usage data collected), and the privacy needs of users (who want as little data collected as possible). This has lead to the development of *cookie files* on one side and *cache busting* on the other. The emerging OPS standard on collecting profile data may be a compromise on what can and will be collected. However, it is not clear how much compliance to this can be expected. Hence, there will be a continual need to develop better instrumentation and data collection techniques, based on whatever is possible and allowable at any point in time.

Portions of Web usage data exist in sources as diverse as *Web server logs, referral logs, registration files,* and *index server logs.* Intelligent integration and correlation of information from these diverse sources can reveal usage information which may not be evident from any one of them. Techniques from data integration [33] should be examined for this purpose.

Web usage data collected in various logs is at a very fine granularity. Therefore, while it has the advantage of being extremely general and fairly detailed, it also has the corresponding drawback that it cannot be analyzed directly, since the analysis may start focusing on micro trends rather than on the macro trends. On the other hand, the issue of whether a trend is micro or macro depends on the purpose of a specific analysis. Hence, we believe there is a need to group individual data collection events into groups, called *Web transactions* [12], before feeding it to the mining system. While [11, 12, 36] have proposed techniques to do so, more attention needs to be given to this issue.

6.2 The Mining Process

The key component of Web mining is the mining process itself. As discussed in this paper, Web mining has adapted techniques from the field of data mining, databases, and information retrieval, as well as developing some techniques of its own, e.g. *path analysis.* A lot of work still remains to be done in adapting known mining techniques as well as developing new ones.

Web usage mining studies reported to date have mined for *association rules, temporal sequences, clusters*, and *path expressions.* As the manner in which the Web is used continues to expand, there is a continual need to figure out new kinds of knowledge about user behavior that needs to be mined.

The quality of a mining algorithm can be measured both in terms of how *effective* it is in mining for knowledge and how *efficient* it is in computational terms. There will always be a need to improve the performance of mining algorithms along both these dimensions.

Usage data collection on the Web is incremental in nature. Hence, there is a need to develop mining algorithms that take as input the existing data, mined knowledge, and the new data, and develop a new model in an efficient manner.

Usage data collection on the Web is also distributed by its very nature. If all the data were to be integrated before mining, a lot of valuable information could be extracted. However, an approach of collecting data from all possible server logs is both non-scalable and impractical. Hence, there needs to be an approach where knowledge mined from various logs can be integrated together into a more comprehensive model.

6.3 Analysis of Mined Knowledge

The output of knowledge mining algorithms is often not in a form suitable for direct human consumption, and hence there is a need to develop techniques and tools for helping an analyst better assimilate it. Issues that need to be addressed in this area include usage analysis tools and interpretation of mined knowledge.

There is a need to develop tools which incorporate statistical methods, visualization, and human factors to help better understand the mined knowledge. Section 4 provided a survey of the current literature in this area.

One of the open issues in data mining, in general, and Web mining, in particular, is the creation of intelligent tools that can assist in the interpretation of mined knowledge. Clearly, these tools need to have specific knowledge about the particular problem domain to do any more than filtering based on statistical attributes of the discovered rules or patterns. In Web mining, for example, intelligent agents could be developed that based on discovered access patterns, the topology of the Web locality, and certain heuristics derived from user behavior models, could give recommendations about changing the physical link structure of a particular site.

7 Conclusion

The term *Web mining* has been used to refer to techniques that encompass a broad range of issues. However, while meaningful and attractive, this very broadness has caused Web mining to mean different things to different people [21, 36], and there is a need to develop a common vocabulary. Towards this goal we proposed a definition of Web mining, and developed a taxonomy of the various ongoing efforts related to it. Next, we presented a survey of the research in this area and concentrated on Web usage mining. We provided a detailed survey of the efforts in this area, even though the survey is short because of the area's newness. We provided a general architecture of a system to do Web usage mining, and identified the issues and problems in this area that require further research and development.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, Santiago, Chile, 1994.

[2] S. Agrawal, R. Agrawal, P.M. Deshpande, A. Gupta, J. Naughton, R. Ramakrishna, and S. Sarawagi. On the computation of multidimensional aggregates. In *Proc. of the 22nd VLDB Conference*, pages 506–521, Mumbai, India, 1996.

[3] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. Webwatcher: A learning apprentice for the world wide web. In *Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments.* 1995.

[4] M. Balabanovic, Yoav Shoham, and Y. Yun. An adaptive agent for automated web browsing. *Journal of Visual Communication and Image Representation*, 6(4), 1995.

[5] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G Zweig. Syntactic clustering of the web. In *Proc. of 6th International World Wide Web Conference*, 1997.

[6] C. M. Brown, B. B. Danzig, D. Hardy, U. Manber, and M. F. Schwartz. The harvest information discovery and access system. In *Proc. 2nd International World Wide Web Conference*, 1994.

[7] P. Buneman, S. Davidson, G. Hillebrand, and D. Suciu. A query language and optimization techniques for unstructured data. In *Proc. of 1996 ACM-SIGMOD Int. Conf. on Management of Data*, 1996.

[8] P. Buneman, S. Davidson, and D. Suciu. Programming constrcuts for unstructured data. In *Proceedings of ICDT'95, Gubbio, Italy*, 1995.

[9] C. Chang and C. Hsu. Customizable multi-engine search tool with clustering. In *Proc. of 6th International World Wide Web Conference*, 1997.

[10] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Irland, Y. Papakonstantinou, J. Ulman, and J. Widom. The tsimmis project: Integration of heterogenous information sources. In *Proc. IPSJ Conference*, Tokyo, 1994.

[11] M.S. Chen, J.S. Park, and P.S. Yu. Data mining for path traversal patterns in a web environment. In *Proceedings of the 16th International Conference on Distributed Computing Systems*, pages 385–392, 1996.

[12] R. Cooley, B. Mobasher, and J. Srivastava. Grouping web page references into transactions for mining world wide web browsing patterns. Technical Report TR 97-021, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.

[13] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. Technical Report TR 97-027, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.

[14] R. B. Doorenbos, O. Etzioni, and D. S. Weld. A scalable comparison shopping agent for the world wide web. Technical Report 96-01-03, University of Washington, Dept. of Computer Science and Engineering, 1996.

[15] C. Dyreson. Using an incomplete data cube as a summary data sieve. *Bulletin of the IEEE Technical Committee on Data Engineering*, pages 19–26, March 1997.

[16] e.g. Software Inc. Webtrends. *http://www.webtrends.com*, 1995.

[17] W. B. Frakes and R. Baeza-Yates. *Information Retrieval Data Structures and Algorithms.* Prentice Hall, Englewood Cliffs, NJ, 1992.

[18] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. In *IEEE 12th International Conference on Data Engineering*, pages 152–159, 1996.

[19] K. Hammond, R. Burke, C. Martin, and S. Lytinen. Faq-finder: A case-based approach to knowledge navigation. In *Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments.* AAAI Press, 1995.

[20] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. In *IEEE Transactions on Knowledge and Data Eng.*, volume 5, pages 29–40, 1993.

[21] J. Han, Y. Fu, W. Wang, K. Koperski, and O. Zaiane. Dmql: A data mining query language for relational databases. In *SIGMOD'96 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada, 1996.

[22] V. Harinarayan, A. Rajaraman, and J.D. Ullman. Implementing data cubes efficiently. In *Proc. of 1996 ACM-SIGMOD Int. Conf. on Management of Data*, pages 311–322, Montreal, Canada, 1996.

[23] Open Market Inc. Open market web reporter. *http://www.openmarket.com*, 1996.

[24] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons, 1990.

[25] I. Khosla, B. Kuhn, and N. Soparkar. Database search using informatiuon mining. In *Proc. of 1996 ACM-SIGMOD Int. Conf. on Management of Data*, 1996.

[26] R. King and M. Novak. Supporting information infrastructure for distributed, heterogeneous knowledge discovery. In *Proc. SIGMOD 96 Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Canada, 1996.

[27] T. Kirk, A. Y. Levy, Y. Sagiv, and D. Srivastava. The information manifold. In *Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments.* AAAI Press, 1995.

[28] D. Konopnicki and O. Shmueli. W3qs: A query system for the world wide web. In *Proc. of the 21th VLDB Conference*, pages 54–65, Zurich, 1995.

[29] M. Koster. Aliweb - archie-like indexing in the web. In *Proc. 1st International Conference on the World Wide Web*, pages 91–100, May 1994.

[30] C. Kwok and D. Weld. Planning to gather information. In *Proc. 14th National Conference on AI*, 1996.

[31] L. Lakshmanan, F. Sadri, and I. N. Subramanian. A declarative language for querying and restructuring the web. In *Proc. 6th International Workshop on Research Issues in Data Engineering: Interoperability of Nontraditional Database Systems (RIDE-NDS'96)*, 1996.

[32] 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.

[33] E. Lim, S.Y. Hwang, J. Srivastava, D. Clements, and M. Ganesh. Myriad: design and implementaion of federated database prototype. *Software - Practive & Experience*, 25(5):533–562, 1995.

[34] Y. S. Maarek and I.Z. Ben Shaul. Automatically organizing bookmarks per content. In *Proc. of 5th International World Wide Web Conference*, 1996.

[35] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovering frequent episodes in sequences. In *Proc. of the First Int'l Conference on Knowledge Discovery and Data Mining*, pages 210–215, Montreal, Quebec, 1995.

[36] B. Mobasher, N. Jain, E. Han, and J. Srivastava. Web mining: Pattern discovery from world wide web transactions. Technical Report TR 96-050, University of Minnesota, Dept. of Computer Science, Minneapolis, 1996.

[37] net.Genesis. net.analysis desktop. *http://www.netgen.com*, 1996.

[38] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In *Proc. of the 20th VLDB Conference*, pages 144–155, Santiago, Chile, 1994.

[39] K. A. Oostendorp, W. F. Punch, and R. W. Wiggins. A tool for individualizing the web. In *Proc. 2nd International World Wide Web Conference*, 1994.

[40] P. Merialdo P. Atzeni, G. Mecca. Semistructured and structured data in the web: Going back and forth. In *Proceedings of the Workshop on the Management of Semistructured Data (in conjunction with ACM SIGMOD)*, 1997.

[41] M. Pazzani, J. Muramatsu, and D. Billsus. Syskill & webert: Identifying interesting web sites. In *Proc. AAAI Spring Symposium on Machine Learning in Information Access*, Portland, Oregon, 1996.

[42] M. Perkowitz and O. Etzioni. Category translation: learning to understand information on the internet. In *Proc. 15th International Joint Conference on AI*, pages 930–936, Montral, Canada, 1995.

[43] P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear: Extracting usable structures from the web. In *Proc. of 1996 Conference on Human Factors in Computing Systems (CHI-96)*, Vancouver, British Columbia, Canada, 1996.

[44] J. Pitkow. In search of reliable usage data on the www. In *Sixth International World Wide Web Conference*, pages 451–463, Santa Clara, CA, 1997.

[45] J. Pitkow and Krishna K. Bharat. Webviz: A tool for world-wide web access log analysis. In *First International WWW Conference*, 1994.

[46] D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. Querying semistructured heterogeneous information. In *International Conference on Deductive and Object Oriented Databases*, 1995.

[47] P. Resnik, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proc. of the 1994 Computer Supported Cooperative Work Conference, ACM*, 1994.

[48] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *Proc. of the 21th VLDB Conference*, pages 432–443, Zurich, Switzerland, 1995.

[49] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proc. of 1995 Conference on Human Factors in Computing Systems (CHI-95)*, pages 210–217, 1995.

[50] A. Shukla, P.M. Deshpande, J. Naughton, and K. Ramaswamy. Storage estimation for multidimensional aggregates in the presence of hierarchies. In *Proc. of the 22nd VLDB Conference*, pages 522–531, Mumbai, India, 1996.

[51] E. Spertus. Parasite: mining structural information on the web. In *Proc. of 6th International World Wide Web Conference*, 1997.

[52] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proc. of the Fifth Int'l Conference on Extending Database Technology*, Avignon, France, 1996.

[53] R. Weiss, B. Velez, M. A. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, and D. K. Gifford. Hypursuit: a hierarchical network search engine that exploits content-link hpertexxt clustering. In *Hypertext'96: The Seventh ACM Conference on Hypertext*, 1996.

[54] S.M. Weiss and C. A. Kulikowski. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.* Morgan Kaufmann, San Mateo, CA, 1991.

[55] M. R. Wulfekuhler and W. F. Punch. Finding salient features for personal web page categorization. In *Proc. of 6th International World Wide Web Conference*, 1997.

[56] O. R. Zaiane and J. Han. Resource and knowledge discovery in global information systems: A preliminary design and experiment. In *Proc. of the First Int'l Conference on Knowledge Discovery and Data Mining*, pages 331–336, Montreal, Quebec, 1995.