# WebSIFT: The Web Site Information Filter System

Robert Cooley*, Pang-Ning Tan, Jaideep Srivastava†
{cooley,ptan,srivasta}@cs.umn.edu
Department of Computer Science
University of Minnesota

June 13, 1999

**Abstract**

*Web Usage Mining* is the application of data mining techniques to large Web data repositories in order to extract usage patterns. As with many data mining application domains, the identification of patterns that are considered *interesting* is a problem that must be solved in addition to simply generating them. A necessary step in identifying interesting results is quantifying what is considered uninteresting in order to form a basis for comparison. Several research efforts have relied on manually generated sets of uninteresting rules. However, manual generation of a comprehensive set of evidence about beliefs for a particular domain is impractical in many cases. Generally, domain knowledge can be used to automatically create evidence for or against a set of beliefs. For Web Usage Mining, there are three types of domain information available; *usage, content*, and *structure*. The Web Site Information Filter (WebSIFT) system uses the content and structure information from a Web site in order to identify potentially interesting results from mining usage data. This paper gives a brief overview of the WebSIFT systems and presents examples of interesting frequent itemsets automatically discovered from real Web data.

*Supported by NSF grant EHR-9554517
†Supported by ARL contract DA/DAKF11-98-P-0359

# 1 Introduction and Background

The World Wide Web continues to expand at an amazing rate as a medium for conducting business and disseminating information. Despite evolving standards and technology, the ability to thoroughly analyze the usage of a Web site remains, and will grow as, an important capability for Web administrators. Design of a Web site involves organizing the information on each page and the hypertext links between pages in a way that seems most natural to the site users in order to facilitate their browsing. For small sites, an individual Web designer's intuition along with some straightforward usage statistics may be adequate for predicting and verifying the users' browsing behavior. However, as the size and complexity of a Web site increases, the statistics provided by existing Web log analysis tools [WLA, HLC, FWP] may prove inadequate, and more sophisticated types of analyses will be necessary. *Web Usage Mining*, which is the application of data mining techniques to large Web data repositories, adds powerful techniques to the tools available to a Web site administrator for analyzing Web site usage.

Web Usage Mining techniques developed in [BM98, CMS99, CPY96, SZAS97, SF98, ZXH98, PE98] have been used to discover frequent itemsets, association rules, clusters of similar pages and users, sequential patterns, and to perform path analysis. Several research efforts [NW97, JFM97] have considered usage information in order to perform *Web Content Mining* [CMS97]. In Web Usage Mining, as with many data mining domains, thresholds for values such as *support* and *confidence* are often used to limit the number of discovered rules to a manageable number. However, high thresholds rarely discover any new knowledge and low thresholds usually result in an unmanageable number of rules.

The notion of what makes discovered knowledge interesting has been addressed in [PSM94, ST96, LHC97, PT98]. A common theme among the various criteria for interestingness is the concept of *novelty* or *unexpectedness* of a rule. Results that were previously known by the data analyst are not considered interesting. [PT98] formally defines the unexpectedness of a rule in terms of its deviation from a set of beliefs. [LHC97] has a broader definition of interestingness that includes discovered rules that are not specifically covered by an initial set of beliefs. In other words, a rule that doesn't contradict an existing belief, but points out a relationship that hadn't been considered is also interesting. While both [LHC97] and [PT98] give examples of small sets of manually generated beliefs, neither addresses the problem of automated generation of a realistic belief set from a large amount of data.

The Web Site Information Filter (WebSIFT) system is a Web Usage Mining framework that, in addition to performing preprocessing and knowledge discovery, uses the structure and content information about a Web site to automatically define a belief set. The information filter uses this belief set to identify results that are potentially interesting. This paper gives a brief overview of the WebSIFT system along with some examples of potentially interesting results found from applying the information filter to frequent itemsets found from a University of Minnesota Computer Science department Web server log.

# 2 The WebSIFT System

The WebSIFT system, which is based on the WEBMINER prototype [CMS97], divides the Web Usage Mining process into three main parts, as shown in Figure 1. For a particular Web site, the three server logs - access, referrer, and agent, the HTML files that make up the site, and any optional data such as registration data or remote agent logs provide the input. The preprocessing phase uses the input data to construct a user session file, which is the best estimate of the user's

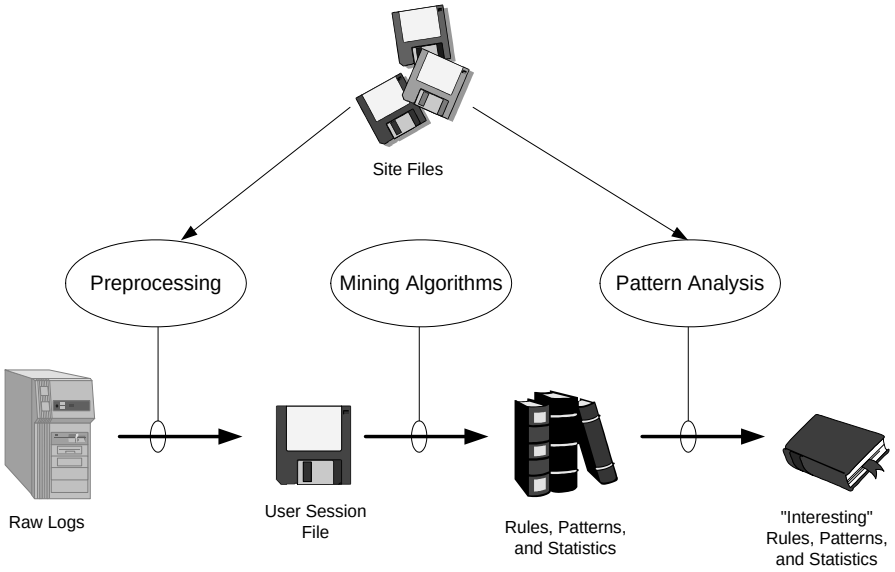


Figure 1: High Level *Web Usage Mining* Process

browsing behavior based on the methods and heuristics discussed in [CMS99]. In addition to being used to derive a site topology, the site files are used to classify the pages of a site. Both the site topology and page classifications are then fed into the *information filter*, which is described in the next section.

The knowledge discovery phase uses existing data mining techniques to generate rules and patterns. Included in this phase is the generation of general usage statistics, such as number of "hits" per page, page most frequently accessed, most common starting page, and average time spent on each page. The discovered information is then fed into various pattern analysis tools. The WebSIFT system, shown in detail in Figure 2, has been implemented using a relational database, procedural SQL, and the Java programming language. Java Database Connectivity (JDBC) drivers are used to interface with the database. Although algorithms have been identified and tested for individual portions of the system, only the generation and filtering of frequent itemsets, association rules, and general statistics is fully automated at this time.

# 3 Information Filtering

For Web Usage Mining, there are two sets of domain knowledge that can provide evidence about beliefs; the content data and the structure data. It is assumed that content and structure data can be used as surrogates for the Web site designer's domain knowledge. Links between pages provide evidence which supports the belief that those pages are related. The strength of the evidence for a set pages being related is proportional to the strength of the topological connection between the set of pages. Evidence based on the site content can also be automatically generated by looking at content similarity, and by calculating the "distance" between pages. Table 1 gives some examples of the types of interesting beliefs that can be identified in the Web Usage Mining domain.

The current implementation of the WebSIFT system uses two different methods to identify interesting results from a list of discovered frequent itemsets. The first is to declare itemsets that contain pages not directly connected to be interesting. This corresponds to a situation where a belief that a set of pages are related has no domain or existing evidence, but there is mined evidence.

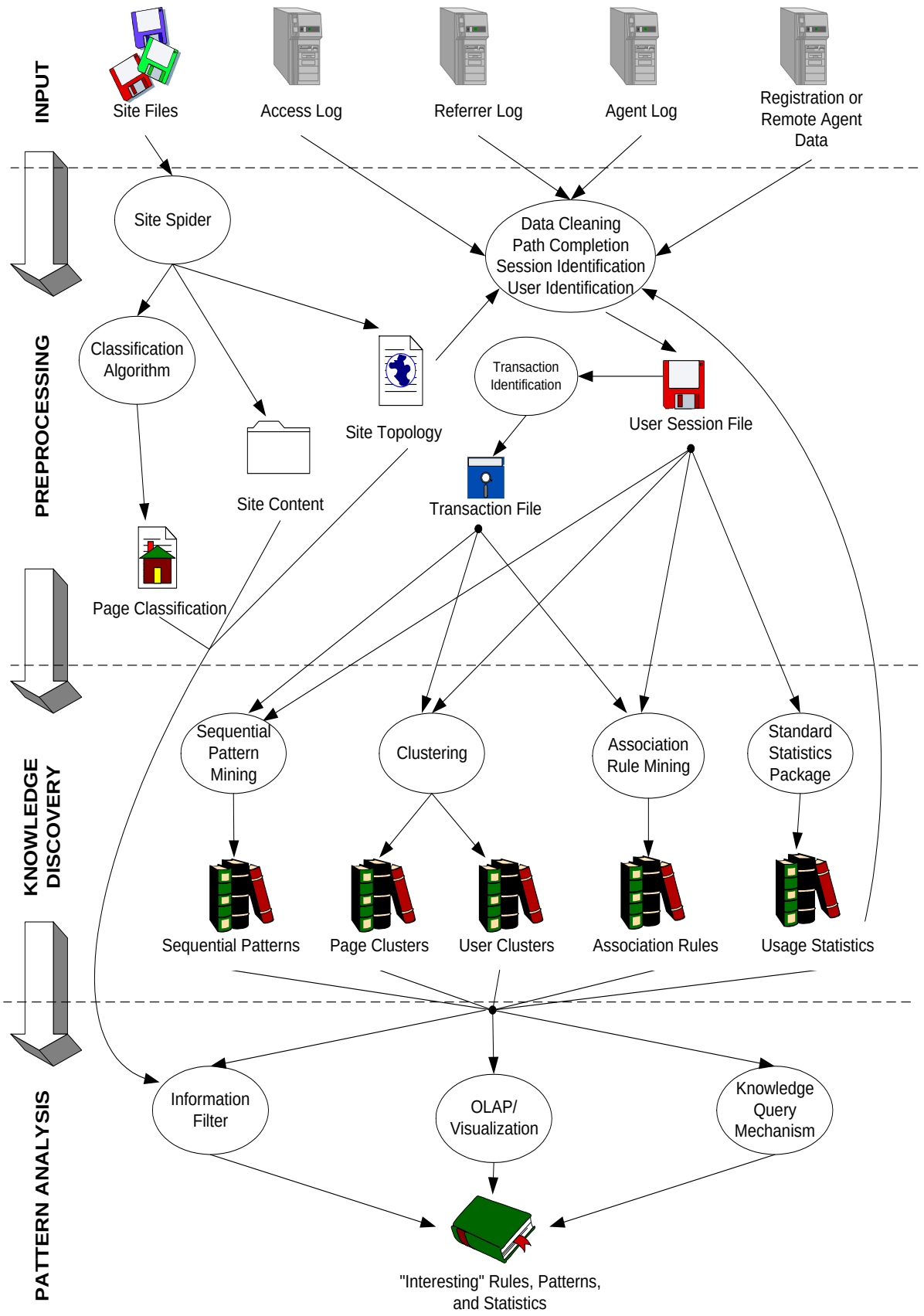


Figure 2: WebSIFT Architecture

Table 1: Examples of Web Usage Information that can be automatically flagged as Interesting

| **Mined Knowledge Source** | **Domain Knowledge Source** | **Interesting Belief Example** |
|---|---|---|
| General Usage Statistics | Site Structure | The head page is not the most common entry point for users |
| General Usage Statistics | Site Content | A page that is designed to provide content is being used as a navigation page |
| Frequent Itemsets | Site Structure | A frequent itemset contains pages that are not directly linked |
| Usage Clusters | Site Content | A usage cluster contains pages from multiple content clusters |

This algorithm is referred to as the Beliefs with Mined Evidence (BME) algorithm in Section 4. In the second approach the absence of certain frequent itemsets is interpreted by the information filter as evidence *against* a belief that pages are related. Pages which have individual support above a threshold, but are not present together in larger frequent itemsets provide mined evidence against the pages being related. If the domain evidence suggests that the pages are related (the pages are linked), the absence of the frequent itemset can be considered interesting. This situation of contradicting domain and mined evidence is handled by the Beliefs with Contradicting Evidence (BCE) algorithm.

# 4 Experimental Evaluation

The experiments described in this section were performed on a Web server log from the University of Minnesota Department of Computer Science and Engineering Web site;
http://www.cs.umn.edu/. The server log collects data in the combined log format, with the access, agent, and referrer data all collected in a single file. The log used spanned eight days in February, 1999. The physical size of the log was 19.3 MB and it consisted of 102,838 entries in its raw form. Once preprocessing was completed, there were 43,158 page views divided among 10,609 user sessions.

A threshold of 0.1% for support was used to generate 693 frequent itemsets, with a maximum set size of six pages. There were 178 unique pages represented in all of the rules. Both the BME and BCE algorithms described in the previous section were run on the frequent itemsets. The BME algorithm resulted in 11 frequent itemsets being declared as potentially interesting, and the BCE algorithm resulted in 10 missing page pairs being declared as potentially interesting. Tables 2 and 3 show the interesting results identified by each algorithm.

Of the frequent itemsets shown in Table 2, the two including the graduate handbook (numbers 10 and 11) are of note because these pages are out-of-date. A page with the 1998-99 graduate handbook exists, and the links from the /grad-info/ page to the older handbooks have been removed. However, since the pages were not actually removed from the site and other pages in the site reference them (or users have old bookmarks), the older handbooks are still accessed. The supports of these itemsets are 0.25% and 0.22% respectively. Had the support threshold been set higher to limit the total number of itemsets discovered, the rules would have been missed.

In Table 3, the fourth pair of pages is of note because the first page functions solely as an entry page to a particular research group's pages. However, the link from the first page is flashing and

Table 2: Interesting frequent itemsets identified by BME algorithm

| # | Mined Support(%) | Related Pages |
|---|---|---|
| 1 | 0.10 | /Research/, /tech_reports/ |
| 2 | 0.10 | /employment/, /newsletter/ |
| 3 | 0.10 | /faculty/, /newsletter/ |
| 4 | 0.10 | /icra99/ICRA99-Index.htm, /icra99/Notice.html, /icra99/TechnProgram.htm, /icra99/advanceprogram2.htm |
| 5 | 0.10 | /new/, /sem-coll/ |
| 6 | 0.10 | /reg-info/98-99_schedule.html, /reg-info/ss1-99.html, /reg-info/ss2-99.html |
| 7 | 0.11 | /Research/Agassiz/, /faculty/ |
| 8 | 0.11 | /icra99/Notice.html, /icra99/best.html |
| 9 | 0.11 | /icra99/Proceeding-Order.htm, /icra99/Registration.htm |
| 10 | 0.22 | /grad-info/, /grad-info/97-98-grad-handbook.html |
| 11 | 0.25 | /grad-info/, /grad-info/96-97-grad-handbook.html |

Table 3: Interesting page pairs identified by BCE algorithm

| # | Web Pages |
|---|---|
| 1 | /Research/Agassiz/agassiz_pubs.html, /Research/Agassiz/agassiz_people.html |
| 2 | /Research/GIMME/tclprop.html, /Research/GIMME/Nsync.html |
| 3 | /Research/airvl/minirob.html, /Research/airvl/loon.html |
| 4 | /Research/mmdbms/home.shtml, /Research/mmdbms/group.html |
| 5 | /newsletter/kumar.html, /newsletter/facop.html |
| 6 | /newsletter/letter.html, /newsletter/facop.html |
| 7 | /newsletter/letter.html, /newsletter/kumar.html |
| 8 | /newsletter/newfac.html, /newsletter/facop.html |
| 9 | /newsletter/newfac.html, /newsletter/kumar.html |
| 10 | /newsletter/newfac.htm, /newsletter/letter.html |

located fairly low on the page. This indicates a design problem since not all of the visitors from the first page are visiting the second.

# 5 Conclusions

This paper has shown how even the simplest use of structure information to represent domain knowledge is highly effective in filtering discovered rules. From a set of almost 700 discovered frequent itemsets, 21 interesting itemsets were identified. Of those 21 interesting itemsets, two identified out-of-date information that needed to be removed from a Web site, and one pointed out an instance of poor page design. Future work will include filtering frequent itemsets, sequential patterns, and clusters discovered from usage data using both structure and content data. In order to extend the simple boolean logic used in the BME and BCE algorithms, probabilities and fuzzy logic will be incorporated into the information filter.

## References


[BM98] Alex Büchner and Maurice D Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. *SIGMOD Record*, 27(4):54–61, 1998.

[CMS97] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Web mining: Information and pattern discovery on the world wide web. In *International Conference on Tools with Artificial Intelligence*, pages 558–567, Newport Beach, 1997. IEEE.

[CMS99] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1), 1999.

[CPY96] M.S. Chen, J.S. Park, and P.S. Yu. Data mining for path traversal patterns in a web environment. In *16th International Conference on Distributed Computing Systems*, pages 385–392, 1996.

[FWP] Funnel web professional. http://www.activeconcepts.com.

[HLC] Hit list commerce. http://www.marketwave.com.

[JFM97] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In *The 15th International Conference on Artificial Intelligence*, Nagoya, Japan, 1997.

[LHC97] Bing Liu, Wynne Hsu, and Shu Chen. Using general impressions to analyze discovered classification rules. In *Third International Conference on Knowledge Discovery and Data Mining*, 1997.

[NW97] D.S.W. Ngu and X. Wu. Sitehelper: A localized agent that helps incremental exploration of the world wide web. In *6th International World Wide Web Conference*, Santa Clara, CA, 1997.

[PE98] Mike Perkowitz and Oren Etzioni. Adaptive web sites: Automaticlly synthesizing web pages. In *Fifteenth National Conference on Artificial Intelligence*, Madison, WI, 1998.

[PSM94] G. Piatetsky-Shapiro and C. J. Matheus. The interestingness of deviations. In *AAAI-94 Workshop on Knowledge Discovery in Databases*, pages 25–36, 1994.

[PT98] Balaji Padmanabhan and Alexander Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Fourth International Conference on Knowledge Discovery and Data Mining*, pages 94–100, New York, New York, 1998.

[SF98] Myra Spiliopoulou and Lukas C Faulstich. Wum: A web utilization miner. In *EDBT Workshop WebDB98*, Valencia, Spain, 1998. Springer Verlag.

[ST96] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Eng.*, 8(6):970–974, 1996.

[SZAS97] Cyrus Shahabi, Amir M Zarkesh, Jafar Adibi, and Vishal Shah. Knowledge discovery from users web-page navigation. In *Workshop on Research Issues in Data Engineering*, Birmingham, England, 1997.

[WLA] Webtrends log analyzer. http://www.webtrends.com.

[ZXH98] O. R. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying olap and data mining technology on web logs. In *Advances in Digital Libraries*, pages 19–29, Santa Barbara, CA, 1998.