Abbreviations:
ROLAP: Relational OLAP
MOLAP: Multidimensional OLAP
TOOL: Web mining tool
APACHE: Web server
BAZAAR: Information extraction tool
SURFAID: Information extraction tool
SPEEDTRACER: Information extraction tool
ACCESSWATCH: Information extraction tool
HTTP-ANALYZE: Information extraction tool

table

ABSTRACT
Analyzing Web Logs for usage and access trends can not only provide important information to web site developers and administrators, but also help in creating adaptive web sites. While there are many existing tools that generate fixed reports from web logs, they typically do not allow ad-hoc analysis queries. Moreover, such tools cannot discover hidden patterns of access embedded in the access logs. We describe a relational OLAP (ROLAP) approach for creating a web-log warehouse. This is populated both from web logs, as well as the results of mining web logs. We also present a web based ad-hoc tool for analytic queries on the warehouse. We discuss the design criteria that influenced our choice of dimensions, facts and data granularity, and present the results from analyzing and mining the logs.

Keywords
Web Mining, clustering, user interface, ad hoc analysis, web logs.

1. INTRODUCTION
Web mining can be viewed as the extraction of structure from an unlabeled, semi-structured data set containing the characteristics of users/information respectively [2]. The logs kept by web servers provide a classic example of such data. Analyzing and exploring regularities in the behavior of the users accessing a web site can improve system performance, enhance the quality and delivery of Internet information services to the end user, and identify population of potential customers for electronic commerce [1]. Many approaches have been suggested to mine information from web access log records collected from servers [3][4][5][8]. Typically data has to be scrubbed and preprocessed as a precursor to mining. Such data can also be stored in a data warehouse and be amenable to OLAP like queries. The objective of this paper is to outline a design used to create a ROALP warehouse from raw and mined web log. We also describe the web based analysis tool we have created that helps users in querying the warehouse in an ad-hoc manner.

Of late, new tools promising to apply data warehousing and mining techniques on web logs have entered the market. These include SurfAid & SpeedTracer from IBM, Bazaar Analyzer etc.
In designing the access log warehouse, we included each log entry as a record in the warehouse, which was created in an Oracle 8 database using star schema. Before loading log entries into the warehouse, we reformatted them to conform with the formats that Oracle SQL Loader accepts. This was done using a Perl script we created by modifying Follow [11]. Log structure was assumed to be that prescribed by the Common Log Format, \{<domain> <identity> <authorization> <date> <method> <URL> <protocol> <status> <size returned>\} and any entry not conforming to this structure was discarded. We also disregarded the identity and authorization attributes in the web log since they were not being logged. We have included log entries that resulted in errors or redirects in our analysis.

For the mining part, we filtered out other unwanted entries like record accesses to image files that were embedded in the web pages whose 'hit' had already been logged. For the preprocessing phase, we created a file listing all the sessions (described in the next section) obtained from the logs. The session file consists of Session number and the Domain of that session. A URL file was created that contained all the URLs in the web log that was analyzed and a unique number associated with the URL. Once the sessions were clustered as described in later sections, a cluster file listed each session and the cluster that it belongs to. Using the three files as our input, we proceeded to populate the warehouse for the mined data.

3. DATA WAREHOUSE ARCHITECTURE

In a data warehouse Facts defined along with certain attributes (called dimensions) are the core data elements being analyzed. In this system, the web log entries are facts. Nasraoui et al have suggested the following dimensions [4]: Domain name of the request, date of the request, method of the request, URL of the page accessed, Result of the request, and Number of bytes transmitted.

The fact tables in the warehouse were directly created from the web logs. The web log data is analogous to the transactional data in an OLTP system and the main fact table was populated from it. This table contained all the log entries that were ever made in our web logs (Figure 2(b)). While such a table can answer any analysis query, aggregation queries (find how many queries resulted in a redirect) may take time. Based on some expected aggregation queries, other fact tables were created that contained data related to a specific aggregation criterion. For example, we created separate fact tables for each month of the log. If a user specified the month in the 'date accessed' field (see figure 4(a)), then the fact table corresponding to the month queried would be accessed.

As mentioned earlier, our web analysis does not build on an existing transactional database, since the web server uses a flat file to store its transactional data (accesses) without any normalization. Figure 2(a) illustrates the table structure for a corresponding normalized, relational database. It has "Dimension" tables for data such as Request status codes, request methods, Web protocols etc. The arrows indicate the relationships between dimensions. The dimension data is typically static and dependent on the domain. It also helps map "codes" relating to the HTTP protocol to their descriptions in an efficient manner.

In addition to the fact tables, a normalized schema was added to the warehouse to populate the mining results so as to facilitate on line analysis. Figure 3 represents the Entity-Relationship diagram of this database which consists of three tables. The URL table stores the URL description along with the Unique ID (URL_NO) that is generated by the program. Session Table contains data pertaining to the Domain that identifies the session, the cluster number to which the session belongs and the frequency of the domain. Session-No is the unique identifier for each session. We know that many URLs (say N) can be accessed in a single session and also one URL can belong in multiple sessions (say N). To incorporate the N:N relationship between URLs and Sessions, we split the relationship into a 1:N relation by introducing a new table SES_URL table. This table contains the primary identifiers of both the Session and URL tables and is populated from the output from the clustering experiment.
4. MINING WEB LOGS

To mine knowledge from the web logs we used two data mining techniques, namely Clustering and Association Rules generation. For the data mining experiments, we collected a variety of logs and ran the association rules generator and clustering algorithm on it. Before mining information from the logs we created sessions from the logs. For discovering association rules, we used SGI's MineSet that implements a variation of the Apriori algorithm [9]. For Clustering we used the Fuzzy C-medoids algorithm developed by Krishnapuram et al [7]. This approach allows us to capture a graded (as opposed to binary) notion of similarity between sessions.

4.1 Session Generation

We now briefly describe the technique [5] to generate user sessions, after pre-processing the log files. Since web servers do not typically log usernames (unless identd is used), this technique defines a user session as accesses from the same IP address such that the duration of time elapsed between any two consecutive accesses in the session is within a prespecified threshold (in our case 45 minutes). This distinguishes it from other approaches such as [8] which assume that a user's identity is known. Each URL in the site is assigned a unique number \( j \) \( \in \{1, ... ,|U|\} \), where \( |U| \) is the total number of valid URLs. Analogous to [5], the \( i^{th} \) user session is encoded as an \(|U|-dimensional binary attribute vector \( s^{(i)} \) with the property

\[
s^{(i)}_j = \begin{cases} 1 & \text{if user accessed the } j^{th} \text{ URL in ith session} \\ 0 & \text{otherwise} \end{cases}
\]

This schema will match one user's sessions into multiple user sessions. However, Krishnapuram et al argue that this will not affect the study if one is looking for 'typical user profiles'. Majority of a user's sessions follow a similar pattern, then clearly no difference is made. If the same user displays a few (different) access patterns on a site, this notion of multiple user sessions enables one to capture this behaviour.

4.2 Association Rules

To generate the association rules for the web logs, we first generated the sessions and URL files. MineSet requires two input files for generating the association rules - a schema file and a data file. The schema file describes the structure/fields of the data file. The data file consisted of the binary session vectors. Each session can be thought of as a transaction and each URL as an item, in the association rule context. Thus a session [10110] would imply that URLs 1,3 and 4 were visited in this session and URLs 2 and 5 were not. The schema file was simply the URL numbers and was generated using a Perl program. The association file generated by MineSet was cleaned before the results could be analyzed.

4.3 Clustering

Before generating clusters for our sessions, we have to define a measure of similarity between our sessions. The similarity measure that we used for our experiments has been proposed by Nosraoui et al [6]. They have defined a dissimilarity measure between all session pairs (i.e., the relation matrix) prior to the clustering process. Note that since sessions are not object data, a distance measure in the sense of Minkowski norms is not automatically available. The following paragraph briefly explains the similarity measure used.

Consider two user-sessions \( s^{(k)} \) and \( s^{(m)} \). First consider the case where URLs accessed in the sessions are considered independent of the structure of the site. Then, the cosine of the angle between \( s^{(k)} \) and \( s^{(m)} \) (called \( M_1 \)) is a measure of similarity. \( M_1 \) simply measures the number of identical URLs accessed during the two sessions relative to the number of URLs accessed in both sessions and completely ignores the hierarchical organization of the web site. This adversely affects its ability to capture correct profiles. Another similarity measure [6] takes into account the site structure. The web site is modeled as a tree; essentially the directory structure rooted at the server's document root, with links (such as redirects and aliases) explicitly brought in. Taking into account the syntactic representation of two URLs, their similarity is assessed by comparing the location of their corresponding nodes on the tree. This is done by comparing the paths from the root of the tree to the two nodes. The "syntactic" similarity between the \( i^{th} \) and \( j^{th} \) URLs is

\[
S_u(i,j) = \min \left( 1, \frac{|p_i \cap p_j|}{\max(1, \max(p_i, p_j) - 1)} \right)
\]

where \( p_i \) denotes the path traversed from the root node to the node corresponding to the \( i^{th} \) URL, and \( |p_i| \) indicates the length of this path or the number of edges included in the path. This similarity, which lies in \([0,1]\), measures the number of identical URLs accessed during the two sessions relative to the number of URLs accessed in both sessions.

Due to similarity underestimation by both \( M_1 \) and \( M_2 \) in specific cases, they propose \( m = \max(M_1, M_2) \) as a similarity measure. This similarity is mapped to the dissimilarity measure \( d = (1-m)^2 \).

4.3.1 Evaluating the results

We interpret the results of applying the clustering algorithm on the user session data by using the following quantitative measures. The clustering algorithm [7] assigns user sessions to the closest clusters based on the similarity measures. This creates \( C \) clusters \( \Pi = \{s^{(k)} \in S | d_{ik} \leq d_{ij} \\forall j \neq i\} \), for \( 1 \leq i \leq C \). The sessions in cluster \( \Pi_i \) are summarized in a typical session "profile" vector \( p_i = (p_{i1}, ... , p_{i|U|}) \). The components of \( p_i \) represent the degree with which
each URL belongs to a session. This is computed as:

\[ D_{ij} = p(s_j^{(k)} = 1 \mid s_j^{(k)} \in X_i) = |X_{ij}| / |X_i| \]

where \( X_{ij} = \{ s_j^{(k)} \in X_i \mid s_j^{(k)} > 0 \} \).

In other words, it measures the fraction of the clusters in the session which included this URL. This helps us recognize those URL which form the "core" of the profile. For some URLs the degree is greater than 1. This is because we grouped together all the URLs that belonged to the same directory. Besides summarizing profiles, the components of the profile vector can be used to recognize an invalid profile, which has no strong or frequent access pattern.

5. WEB QUERY TOOL AND PRELIMINARY RESULTS

We have created a web based thin client as the user interface to the system. By using a web browser the user interface becomes platform independent unlike proprietary tools such as Discoverer. The interface is exported as an HTML forms document from the web server that interacts with the Warehouse using CGI and SQLPLUS. The web interface was created using Perl CGI scripts. The user can enter (or select) a value into the fields, and the system automatically constructs an appropriate SQL query. The generated query is piped to SQLPLUS, which runs the query saved in the file and sends the result back to the interface. The result page shows the query that was used by the database along with the retrieved data. It also shows a count of the records retrieved. The web interface is very simple and user friendly. The users do not need to be aware of the underlying database architecture or SQL to use this tool optimally.

We mined a large number of logs for our experiments, like the UMBC Server logs, UMBC CS Department Server logs, logs of some research groups, individual home page logs etc. While analyzing logs for UMBC’s main server for instance, we obtained traversal paths of users accessing the Semester Schedule. Another traversal path corresponded to users accessing the Library web pages to check the availability of books etc. Then there were traversal paths obtained for generic or new users who traversed through most of the links on the main page etc.

5.1 Ad-hoc Analysis

The web interface allows users to perform ad-hoc analysis of both the web log warehouse and the mining results. By integrating both the components on a single tool, we have made it easier to analyze the access trends. A large number of analysis queries were given to the system and it produced correct results. We present some examples next.

Figures 4 and 5 show the web based query interface of the log analysis part of our system to the warehouse. The interface provides pull down menus and fill in fields that allow rollup, drill down and slicing queries. For example, for the Status dimension, users can query for all status values, or can drill down and query for log entries whose status values fall in HTTP response ranges such as 100s, 200s etc. The user can also query for individual status codes like 200 (success) or 404(error). This system can also handle HTTP 1.1 error codes and Methods. By default, the attributes have ‘%’ in their fields, which is the wildcard character that aids in ad-hoc queries. E.g. If we want to query on the frequency of users accessing our web site from educational institutions, then we would enter ‘%.edu’ on the Domain field of our interface. Or by entering ‘%Jun 1998’ on the date field we can retrieve all records that were logged in June 1998. Figure 4 illustrates such an example. Thus arbitrary OLAP queries can be performed interactively by the user.

The Users who are aware of the database structure have additional flexibility of writing on-the-fly SQL queries to do arbitrary analyses (Figure 5). This feature of our tool also allows dicing queries (e.g. 'select * from whole_log_fact1 where (doc_size>=200 and doc_size < 1000); '). This will dice in the size dimension and retrieve all accesses where the size of the returned URL was between 200 and 1000 bytes.

Figures 6–8 below show the web interface to our mining results data. The interface provides pull down menus and fill in fields that allow users to query the traversal paths, sessions and the URLs for each application log. For example, if we want to query on the traversal path of users that originate from educational institutions, then we would enter ‘%.edu’ on the Domain field of our interface. Or by entering ‘0’ on the Cluster Number field we can retrieve all URLs that belong to the sessions of that cluster.
Currently, we have populated the fact table directly from the web logs. We can augment our system by loading the data into the transactional database, and then populate the fact tables from it. Stored procedures or triggers could be written to populate specific fact tables with web log as they get added, or in a batch mode. By creating such a transactional database, the pre-processing step will be automated to a large extent, and more dynamic "monitoring" can be done by the system automatically. Features like alerts and warnings can be easily incorporated in this architecture. For example, access patterns that denote security violations would be identified as they happen by the system (e.g. the abuse of the CGI phf script, or attempted buffer overflow attacks). If a security violation is taking place, the site administrator can be alerted or the web site could be made inaccessible. This proactive feature of our architecture can aid in more efficient web site administration. However, this requires modifying the logging module in the web server to store access logs directly into the database.
Finally, our experimentation with the web mining were done with single (most reasonable) values for many of the tunable parameters, such as the time delta involved in sessionizing logs, confidence and support for associations, initializing of the medoids in clustering etc. Clearly, it would be interesting to experiment with and see the effect of these parameters on the eventual results.

7. ACKNOWLEDGMENTS
This work was partially supported by cooperative NSF awards (IIS 9801711 & IIS 980089) to Joshi and Krishnapuram respectively.

8. REFERENCES
[10]'A Listing of Access Log Analyzers', http://www.uu.se/Software/Analyzers/Access-analyzers.html