An Architecture for Web Usage Mining

José Roberto de Freitas Boullosa
Geraldo Xexéo
Programa de Engenharia de Sistemas e Computação - COPPE / UFRJ
beto@ufrj.br, xexeo@cos.ufrj.br

Abstract

When planning a Web site, designers should have not only a clear understanding of user's profiles and site objectives, but also an asserted knowledge of the way users will browse site pages. Analysis of a site visitors' behavior is a powerful tool that can be used to gather invaluable hints about how well the site is reaching its goals. Such analysis involves transformation and interpretation of Web server log records, in order to find hidden, implicit and previously unknown usage patterns, through the use of data mining and knowledge discovery tools and techniques.
This work proposes an architecture for Web usage mining, such that it can be used as a basis for development, testing and implementation of new Web usage mining methods and algorithms. Furthermore, it shows how this architecture can be useful for a Web designer that intends to build sites and pages that adapt themselves automatically according to user's needs.

1. Introduction

The World Wide Web, since its creation in 1989, has grown to become a very large and complex repository of extremely rich and diverse information, accessed minute after minute by users from all regions of the globe. Every Web page access is part of a continuous and overwhelming amount of user clicks, growing uninterruptly, day after day, a non-stop flow often called the "Web clickstream" (KIMBALL & MERZ, 2000).

Discovery of Web usage patterns and understanding of motivations hidden beneath user navigation have become, in recent years, some of the main goals to a legion of researchers in fields ranging from computer networks to databases, including psychology, artificial intelligence and other areas.

One of the basis for this discovering and understanding is the analysis / interpretation of millions of entries recorded in Web servers log files. These log files, together, form, in an unstructured, disorganized (and often inaccurate) way, a picture of the chain of page clicks. "Often inaccurate" because, due to particularities involved in Web navigation process (e.g., presence of proxy servers and use of cache, that cause a lot of "gaps" and "spaces" to appear in the log files), the task of finding with **total confidence** which pages were effectively visited during a certain period of time becomes virtually impossible.

Web usage mining is the field devoted to this kind of discovering. Through investigation of the "clickstream", it tries not only to recover the steps taken by Web users, but also (and above all) to find the patterns that come out from this recovering. After a pre-

processing and transformation phase, one is able to apply data mining techniques to log files data, in an attempt to efficiently achieve this goal.

Yet, many of the spaces found in sequences of log files records probably will never be confidently filled. Thus, Web usage mining activity can be seen as a kind of "Web archaeology", since it is always trying to make predictions and deductions based, at one side, on effectively available observations and, at the other side, on assumptions made in gap elimination process.

There are a lot of commercial tools and programs intended to extract and analyze data recorded in Web server log files. Generally, they focus in trying to find out statistical information, as, for instance, "which were most accessed pages during a certain time interval". Yet, most of these tools are not suitable for the task of discovering more complex patterns, like, for example, "which are the most popular sequences of pages at one particular site". Furthermore, these tools don't have an open architecture, offering very poor customization features.

At the other hand, non-commercial, academic tools and projects proposed for this task do offer more complex pattern recognition options. But they also have their own drawbacks, mainly, the fact that most of them are built having in mind one specific problem, method or mining task.

In this work, we study the main aspects involved in Web usage mining, identifying its weak and strong points. We also present a general architecture called MineraWeb, aimed to develop, test and execute Web usage mining tools, methods and algorithms. In section 2, Web Usage mining is depicted. It's shown how one can divide Web data mining in two different activities: Web usage mining and Web content mining. Then, Web usage mining is described in detail. In section 3, MineraWeb architecture is described. Section 4 is devoted to work conclusions.

2. Web Usage Mining

Web usage mining, is, in fact, part of a broader field: Web mining. Web mining is, for its turn, one of the applications of data mining. That's why, to understand the former two concepts, one must understand the latter.

Data Mining & Knowledge Discovery in Databases

BERRY & LINOFF (1997) argue that data mining would be the automatic or semi-automatic task of exploring and analysing large sets of data, trying to discover significant rules and patterns. MANNILA (1997) takes the problem of data mining in these terms: given a data set *d* and a class *P* of patterns or sentences describing properties of *d*, one could determine if a pattern $p \hat{I} P$ is interesting and occurs enough times in *d*. Thus, the data mining task would be discovering of set *PI* of patterns such as

$$PI(d, P) = \{ p \hat{I} P \mid p \text{ is interesting and occurs in } d \text{ enough times} \}$$

The term "data mining" is many times referred as a synonym for "database mining" and "knowledge discovery in databases" or KDD. In fact, the expression "knowledge discovery in databases" was first used by PIATESTKY-SHAPIRO (2000), when he organized the first "Workshop in Knowledge Discovery in Databases" in Detroit, 1989. Curiously enough, he didn't use the expression "data mining", already largely used by the database community, and justified his choice arguing that "data mining" was "unsexy" and "mining" was "unglamorous". Besides, the expression was used in a pejorative way by some statisticians. that criticized the field.

FAYYAD *et al.* (1996) consider "knowledge discovery in databases" to be a broader field that includes data mining. They define KDD as the non-trivial extraction of potentially useful, implicit and previously unknown information from data. This extraction would be brought to life through some steps, one of them being data mining itself: 1) definition of domains and goals of KDD process; 2) creation of a data set from available data sources; 3) data pre-processing; 4) data transformation; 5) ***data mining***: this step would involve techniques and algorithms that effectively produce the knowledge; 6) analysis and interpretation of the results.

The data mining field uses some tools and techniques in order to achieve its goals of finding patterns and rules from data (BORGES & LEVENE, 1998). These techniques can use different models, including classification, prediction, clustering and temporal series. Techniques include association rules generation, sequence analysis, clustering, classification, memory-based reasoning, decision trees, rule induction, neural networks, genetic algorithms, among others.

Web Mining

The methods of data mining, when applied to the domain of World Wide Web, can help to find out important information, either to ordinary users, or to Web developers. This kind of application of data mining was quickly named Web mining (COOLEY et al, 1997).

Web mining research is always based in some Web navigation model. A navigation model tries to predict and explain how and why users visit Web pages in such different ways. The model also describes which are those ways, trying to classify different types of pages according to them.

CATLEDGE & PITKOW (1995) say that the Web is a kind of open, highly dynamic and collaborative hypermedia system, a "dynamic information ecology" including two main types of user strategies: search and navigation. COVE & WALSH (1988) add a third strategy, "serendipitous browsing", when the users randomly walks through Web pages. These strategies are not excluding, one user is constantly shifting its focus between them.

Web designers must be aware of these strategies when planning a Web site, since there are different needs associated to each one. There's always the risk of users becoming "lost in cyberspace" (NIELSEN, 1990), when these needs are insufficiently mitigated.

Some researchers (BORGES & LEVENE, 1998, LEVENE & LOIZOU, 1999) use statistical models to represent user navigation. In their works, WWW is considered to be a database of pages, described as a directed graph whose nodes are pages and the arcs are hyperlinks between pages. Through association of states to pages and probabilities to links, one can build Markov chain models to represent the navigation process, since this process has a strong regularity from a statistical point of view.

COOLEY *et al. (*1999) identify 5 different types of Web pages:

i) *head pages* are often the root of a site, generally used as the site entry-point by most users;

ii) *content pages* have a lot texts and graphics, long visitation times, few links;

iii) *navigation pages* have few content, but a lot of links, their visitation times are short;

iv) *look-up pages* have few links and content, short visitation times, are used as ***maximal forward references*** (to be explained later in this section);

v) *user pages* hardly have common characteristics. The pages of a site can be either manually classified by site designer, or automatically by some algorithm, like, for example, C4.5, or, alternatively, through some meta-data schema.

ZAÏANE (1999) distinguishes three different types of Web mining: **Web content mining**, **Web structure mining** and **Web usage mining**. **Web content mining** is an activity directed mainly to Web end users trying to find out relevant information from the contents stored in Web documents. It includes textual data mining, concept-indexed mining and agent-based technology. **Web structure mining** tries to infer knowledge from the structure and organization of a site, its pages topology and hyperlinks between pages. It's directed to Web developers. **Web usage mining** is also directed to Web developers; it looks for useful and relevant information hidden in Web server logs, trying to identify user navigational patterns. COOLEY *et al* (1997) distinguishes only between Web content and Web usage mining, considering that structural analysis is part of Web usage mining.

Web Usage Mining

Web usage mining has two different and complementary aspects: while it is useful for systematically analyze user trends, it's also a powerful tool in designing and modifying a Web site structure. There are two different aspects to understand when analyzing site visitors behavior (SPILIOPOULOU, 1999): a) his interests and information he accesses; b) the way this information is accessed. Web Usage Mining activities are concentrated in the second aspect. Those activities conciliate two different perspectives (COOLEY *et al*,1999): how designers expect the site to be used and the way visitors are effectively using the site.

Web usage mining can be divided in at least three different phases (COOLEY *et al*,1999), namely **data preparation**, **pattern discovering** and **pattern analysis and visualization**.

Data preparation

In the first step, **data preparation**, all Web usage data sources must be integrated, cleaned, filtered and transformed, in such a way that irrelevant information will be thrown away, gaps will be possibly filled and user sessions and transactions will be identified. These data sources include mainly Web server log files, but also agent logs and other interfaces – for example, a system directly implemented into an Internet Service Provider that records every user interaction with the Internet (SHAHABI *et al.* 2001).

The main data source for usage mining, Web server log files are generally stored according to NCSA Common Log File format. Every log entry records the traversal from one page to another, storing user IP number or domain name, time and type of access method

(GET, POST, etc.) and address of the page being accessed. This format was later enlarged (Extended Log Format) to include more fields, such as referrer address (Web page that originated the access).

Filtering and cleaning of log files is very important, because they have a lot of unwanted entries, as, for example, accesses to image or sound files. User and session identification is a hard task, since HTTP protocol tracks neither users nor their sessions (a user session is the whole set of pages accessed during a visit to a site). Cache (local or server-based) adds a complication factor to this identification, since it prevents many user accesses from being recorded in log files. If a user hits a page that already is in cache, this access won't be recorded in the log. Proxy servers are also a problem: users accessing through a proxy server appear to be a single user in the log file, since they'll share a single IP address in the log file.

These problems can be circumvented by some techniques: sending of cookies to identify users, cache busting to prevent use of cache, user explicit registration, utilization of agents and so on. Unfortunately, all these solutions have limitations: cookies and cache busting may be disabled by users; cache disabling may reduce the performance; explicit registration arises a lot of privacy concerns, etc.

Trying to identify sessions is even harder than identifying users, since all users accesses and sessions are recorded together in the logs, sorted in a timely fashion. Some heuristics can be used to identify sessions: using time intervals between consecutive log entries, if two accesses from the same user are separated by an interval longer than a threshold, they're considered to be from different sessions. Many products use a threshold of 30 minutes. Another option is using a time-out to identify the end of a session. To find out spaces caused by use of cache, one can try some kind of algorithm of path completion. COOLEY *et al* (1999) use site structure to help identify user and sessions, filling gaps of unrecorded accesses.

Another activity in data preparation is transaction identification. A transaction is a semantically meaningful subset of accesses inside a session. Many products rely on transactions to find out rules during next step of Web usage mining process. There are two different types of transactions: content transactions and navigation transactions. Content transactions include only content pages, ignoring navigation pages. Navigation transactions include all navigation pages used by the user to reach a certain content page. In the process of

transaction identification, one can use successive steps of grouping and dividing transactions, until a point that the result is considered acceptable. In each step, there are three different identification methods that may be used:

i) **identification by reference length**, based on the premise that time spent in a page ("reference length") is correlated to the fact of this page being a content or reference page;

ii) **identification by maximal forward references,** proposed by CHEN *et al.* (1996), in which a transaction is the sequence of all pages visited until the last page immediately before a *backward reference*. A *backward reference* is a page already visited in the transaction. *Forward references*, on the other side, are the pages still not visited. Thus, a *maximal forward reference* is just the last page visited before the first backward reference. This method assumes that *maximal forward references* are content pages, and the other pages in the transaction are navigation pages; algorithm MF (Maximal Forward) was proposed by Chen to find out these types of transactions, called *maximal forward sequences;*

iii) **identification by time windows** is the simplest method: it divides the sessions in transactions with a certain duration.

Pattern discovering

After pre-processing phase, Data Mining methods and algorithms should be applied to user sessions and transactions identified before. These methods must be sometimes slightly modified to adapt themselves to the particularities of Web data. There are many types of analysis to be performed on this data: simples statistical analysis, traversal path analysis, association rules discovering, sequential patterns finding, clustering and classification of pages or paths, etc.

Statistical analysis looks for simple information, like the number of clicks per page, average page reference length and others. **Path analysis** generates directed graphs, where nodes are pages and arcs are links between pages, in order to find out large reference sequences, common traversed paths, etc. **Rules discovering** finds common rules in the format $A \rightarrow B$, meaning that, when page A (antecedent) is visited in a transaction, page B will also be visited in the same transaction. These rules may have different values of confidence and support. Confidence is the percentage between the number of transactions containing both

items of the rule and the number of transactions containing just the antecedent. Support is the percentage of transactions in which the rule is true.

Sequential patterns finding may show, for example, that a given number of users accessing a page has also made an on-line purchase at another page, inside a week interval. **Classification** and **clustering** techniques can be used to group together not only pages that share common characteristics, but also similar sequences of pages, furthermore allowing comparison of these sequences with user profiles, if available (SU *et al.*, 2001).

Pattern analysis and visualization

After pattern discovering, Web Usage Mining applications must provide means for analysis and visualization of these patterns. Those means include statistical, graphical, visualization and query programs. An analysis tool could graphically show, for example, which are the weak points in a site design, appointing pages that are being really useful for visitors.

Query languages can greatly improve designer's ability of finding useful patterns. MINT, for example, is a language that helps designers specify which are intended interestingness parameters in the process of pattern finding in WUM system (SPILIOPOULOU, 1999).

Utilization of data warehousing and OLAP techniques is another way of improving pattern visualization and analysis, since Web usage data share many common characteristics with data stored in a warehouse. KIMBALL & MERZ (2000) describe an integrated vision of Web usage mining they call "Web datahousing", including OLAP tools and development of data warehousing schemas (both "star" and "snow flakes") for Web data storage.

3. MineraWeb architecture

Available Web usage mining products are often closed-architecture proprietary systems. Both commercial and non-commercial systems don't offer many configuration options. The methods they use, the kind of data sources they analyse, the types of reports they generate are often fixed. An additional problem relies in the fact that, when a Web usage mining researcher intends to test a new mining method or algorithm, he must do a lot of auxiliary work, in order to reach his goal.

For example, he has to prepare raw log server data, performing all cleaning and filtering tasks. Then, he is obliged to identify users, sessions and transactions, preparing data to be analysed, and so on. The problem is then stated in this way: he consumes a lot of his efforts working in accessory tasks, instead of devoting all attention to his main concern, the mining method or algorithm to be tested.

MineraWeb offers a common, modular, opened and scalable environment to support all phases of Web usage mining. It's possible to aggregate, at any time, new algorithms for reading, filtering and pre-processing data from different sources, and methods for discovering and analysing usage patterns from that data. All data is integrated in a relational database, called MineraData. Thus, MineraWeb is an architecture intended to become a basis for testing and development of Web Usage mining techniques, a key factor to researchers working in this field (**Figure 1**).

Furthermore, implementation of MineraWeb modules creates a very useful toolbox for Web designers. It allows them, for example, to build adaptive Web sites, designing pages that change according to users needs. In this work, we've implemented a prototype toolbox made up of some MineraWeb modules.

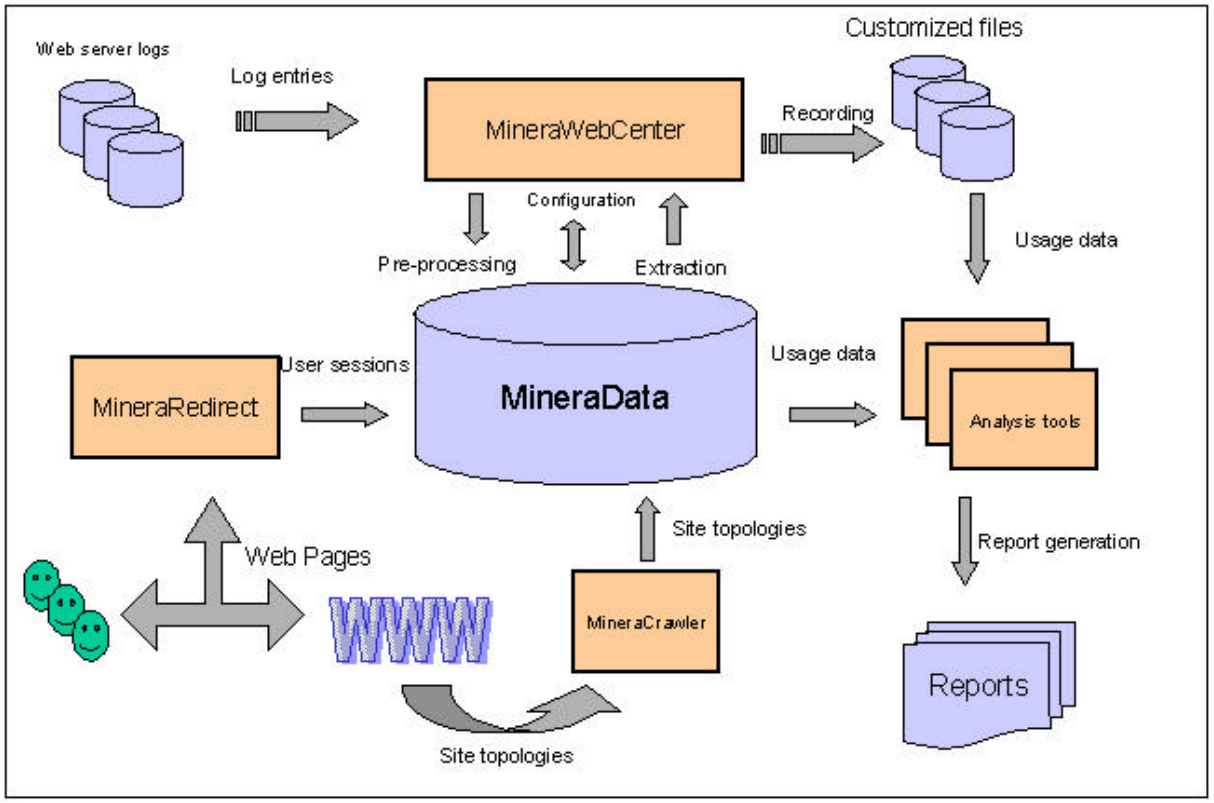

Figure 1: MineraWeb architecture

MineraData

MineraData is the foundation of MineraWeb. A relational data model was planned in order to store key information for Web usage mining. This model is flexible enough to accommodate new data needs that may arise when new algorithms are implemented. MineraData data model provides tables for storing data read from Web server usage logs, after pre-processing steps (**Figure 2**).

Log entries are stored in a main table, VISITS. Every user click becomes a row in VISITS table. The rows keep track of return codes, methods, protocols, server port, time of visit and other useful information about the site clickstream. Most of VISITS columns have a direct correspondence with log entry fields. However, some of their columns, like CODPAGE and other foreign keys, require an amount of investigation work to be filled. Other columns, like VISITLENGTH and CODSESSION, will be filled during data pre-processing phase.

PAGES table represents all Web site pages and their characteristics. Site data is recorded in SITES page. Data about client machines are kept in CLIENTS table. Tables SESSIONS, USERS, TRANSACTIONS and TRANS_ENTRIES will be populated during user, session and transaction identification processes.

In implementing prototype toolbox, MineraData was built on top of SQLServer2000 DBMS in an Windows environment, but, since it's a logical model, it could be easily created in another DBMS.

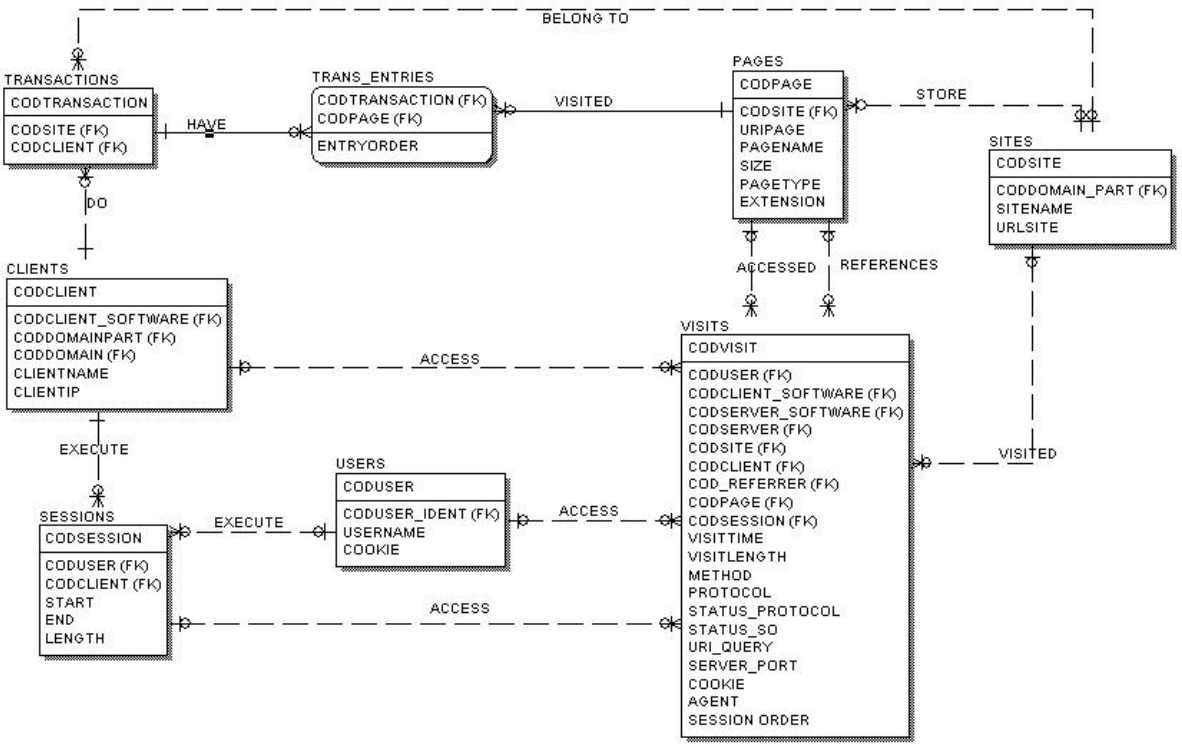

Figure 2: MineraData data model

MineraWebCenter

MineraWebCenter is MineraWeb main module. Its features include toolbox configuration, data loading and filtering, test data generation, user, session and transaction identification, page classification.

Prototype toolbox implemented MineraWebCenter as a C++ application that does data pre-processing in a very flexible way: it uses server-side database stored procedures to filter data, identify users, sessions and transactions. These stored procedures were implemented in TRANSACT-SQL.

Stored procedure IDENTIFY_SESSIONS_TIME uses a method of time windows to identify user sessions. It allows configuration of session maximum length, as proposed by SPILIOPOULOU (1999). Stored procedure IDENTIFY_TRANSACTIONS is used to identify content transactions, using reference length method, plus a time threshold. It assumes that site pages may be divided in content and navigation pages. MineraWebCenter provides an option to classify pages. It identifies a content page based on its reference length.

Filtering in MineraWebCenter is configured through definition of rules similar to SQL integrity rules. For example, rule **@extension NOT IN ('GIF', 'JPG', 'BMP')** would tell filtering process not to load log entries corresponding to image files. MineraWebCenter allows definition of templates for log files. Thus, the user is able to load data from different sources, like, for example, Common Log Format or Extended Log Format files. He must choose, among a set of predefined fields, which ones are present in the data file to be read.

Another interesting feature in MineraWebCenter is the ability of generating log files. Thus, one could use data already stored in MineraData to build a log file directed to some proprietary mining application. Another MineraWebCenter option allows generation of test data to be used in testing mining algorithms.

Analysis and report tools

Beyond pre-processing steps, pattern finding was also implemented in the prototype toolbox. Stored procedure IDENTIFY_RULES was built and used to identify association rules among transactions stored in TRANSACTIONS table. All rules were stored in a set of relational tables later added to the model.

Data can be visualized using third-party tools, like, for example, Microsoft Analysis Services suite. Since log data modeling was built as a data warehouse snow-flake schema, it can be put on a cube and easily visualized. Thus, web designers can, for example, use cube operations, like drill-down and slice & dice to navigate through cube data.

MineraRedirect and MineraCrawler

Rules discovered by IDENTIFY_RULES procedure were used as input for MineraRedirect, a navigation agent aimed to help user navigation, through recommendation of most accessed pages in a set of sites. In toolbox implementation, MineraRedirect is an ISAPI application written in Delphi. It's an interface between user browsers and Web Servers. It catches all user clicks and redirects them to application own site, where MineraWeb is installed. MineraRedirect then stores all user clicks in MineraData database (so, it is also an usage data gathering agent).

It uses stored mined rules to create a framed page with links to "hotter" pages, according to user navigation patterns.

For example, when a user tries to access a given home-page (e.g. Yahoo! home-page), MineraRedirect stores this visit in MineraData. MineraRedirect reads MineraData, and looks

for mined rules that put together Yahoo! home-page and other pages. It then builds a new framed page, one main frame composed of Yahoo! home-page and the other one containing MineraRedirect recommendation links.

Finally, MineraCrawler is an auxiliary tool used to gather site structure, all its pages and links, and store them in MineraData database.

4. Conclusions

Web usage mining is a powerful tool not only to researchers interested in patterns discovering, but also to individuals and organizations involved in design and implementation of Web sites. Web designers have a strong need of detailed information about how users navigate through site pages, in order to take appropriate decisions about structure and topology of the site to be created or modified. Without that information, designers would depend solely on their own assumptions about user expectations and behavior patterns.

MineraWeb proved to be an ideal architecture to Web designers trying to not only get the best from their sites, but also to build adaptive Web sites and pages.

MineraRedirect was used to recommend pages based on usage patterns mined by MineraWeb tools. In a similar fashion, one can build a customizable Web site, with script-based pages that change themselves according to the user, reading MineraData in search of interesting stored rules, paving the way to an entirely adaptive site.

Finally, MineraWeb is a powerful tool for researchers trying to test and develop its own usage mining techniques and algorithms, without having to worry about auxiliary tasks involved in the process, like data cleaning and pre-processing, for example. The prototype toolbox implemented in this work showed how this should be achieved. Implemented modules and tools have a lot of configuration options that allow the user to customize in detail pre-processing tasks. Thus, raw data from Web server logs can be seamlessly loaded into MineraData, leaving the researcher free to focus its efforts in the task of developing and testing mining algorithms.

In the future, we expect to improve the architecture, expanding its data model, allowing insertion of user profiling and demographical information, in order to build a collaborative dynamic Web site. Another possible development is the implementation of different pre-processing and mining tasks, to achieve better performance and provide more sophisticated options to Web designers and Web usage researchers.

5. References

BERRY, A., LINOFF, G, 1997, **Data Mining Techniques: For Marketing, Sales, and Customer Support**, 1 ed., New York, USA, Wiley Computer Publishing.

BORGES, J., LEVENE, M., 1998, "Mining Association Rules in Hypertext Databases", In: **Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD98)**, pp. 149-153, Menlo Park, CA, USA, Aug.

CATLEDGE, L., PITKOW, J., 1995, "Characterizing Browsing Strategies on the World Wide Web", In: **Proceedings of the 3rd International WWW Conference**, Darmstad, Germany, Apr.

COOLEY, R. MOBASHER, B., SRIVASTAVA, J., 1997, "Web Mining: Information and Pattern Discovery on the World Wide Web" (1997), In: **Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI97)**, Newport Beach, CA, USA, Nov.

COOLEY, R. MOBASHER, B., SRIVASTAVA, J., 1999, "Data Preparation for Mining World Wide Web Browsing Patterns", **Knowledge and Information Systems** v.1, n.1 (Jan), pp. 5-32.

COVE, J. F., WALSH, B. C., 1988, "Online text retrieval via browsing", **Information Processing and Management**, v.24, n.1, pp. 31-37.

FAYYAD, U., M., PIATETSKY-SHAPIRO, G., SMYTH, P., 1996, "From data mining to knowledge discovery: An overview", In: Fayyad, U. M.,, Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds), **Advances in Knowledge Discovery and Data Mining**, pp. 1-36, Menlo Park, CA, USA, AAAI Press.

KIMBALL, R., MERZ, R., 2000, **The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse,** 1 ed., New York, USA, John Wiley & Sons.

LEVENE, M., LOIZOU, G, 1999, **Computing the entropy of user navigation in the web**, In: Research Note RN/99/42, Department of Computer Science, University College London, London, UK.

MANNILA, H., 1997, "Methods and problems in data mining", In: **Proceedings of the 6th International Conference on Database Theory (ICDT'97)**, LNCS - Lecture Notes in Computer Science, v. 1186, Springer-Verlag, pp. 41-55, Delphi, Greece, Jan.

NIELSEN, J., 1990, **Hypertext and Hypermedia,** 1 ed., Boston, MA, USA, Academic Press.

PIATESTKY-SHAPIRO, G., 2000, "Knowledge Discovery in Databases: 10 years after", **SIGKDD Explorations**, v. 1, n. 2 (Jan), pp. 59-61.

SHAHABI, C., BANAEI-KASHANI, F., FARUQUE, J., 2001, A Reliable, Efficient, and Scalable System for Web Usage Data Acquisition, In: **Proceedings of the WebKDD'2001 Workshop in conjunction with the ACM-SIGKDD 2001**, San Francisco, CA, Aug.

SPILIOPOULOU, M., 1999, "The laborious way from data mining to web mining", **International Journal of Comp. Sys., Sci. & Eng., Special Issue on "Semantics of the Web"**, v. 14 (Mar), pp. 113-126.

SU, Z., YANG, Q., ZHANG, H., et al., 2001, "Correlation-based Document Clustering using Web Logs", In: **Proceedings of the 34nd IEEE Hawai'i International Conference on System Sciences (HICSS-34)**, Hawaii, USA, Jan.

ZAÏANE, O., R., 1999, **Resource and Knowledge Discovery from the Internet and Multimedia Repositories**, Ph.D. thesis, School of Computing Science, Simon Fraser University, Vancouver, BC, Canada. url: http://www.cs.aue.auc.dk/datamining/papers/osmarzaianephd.pdf