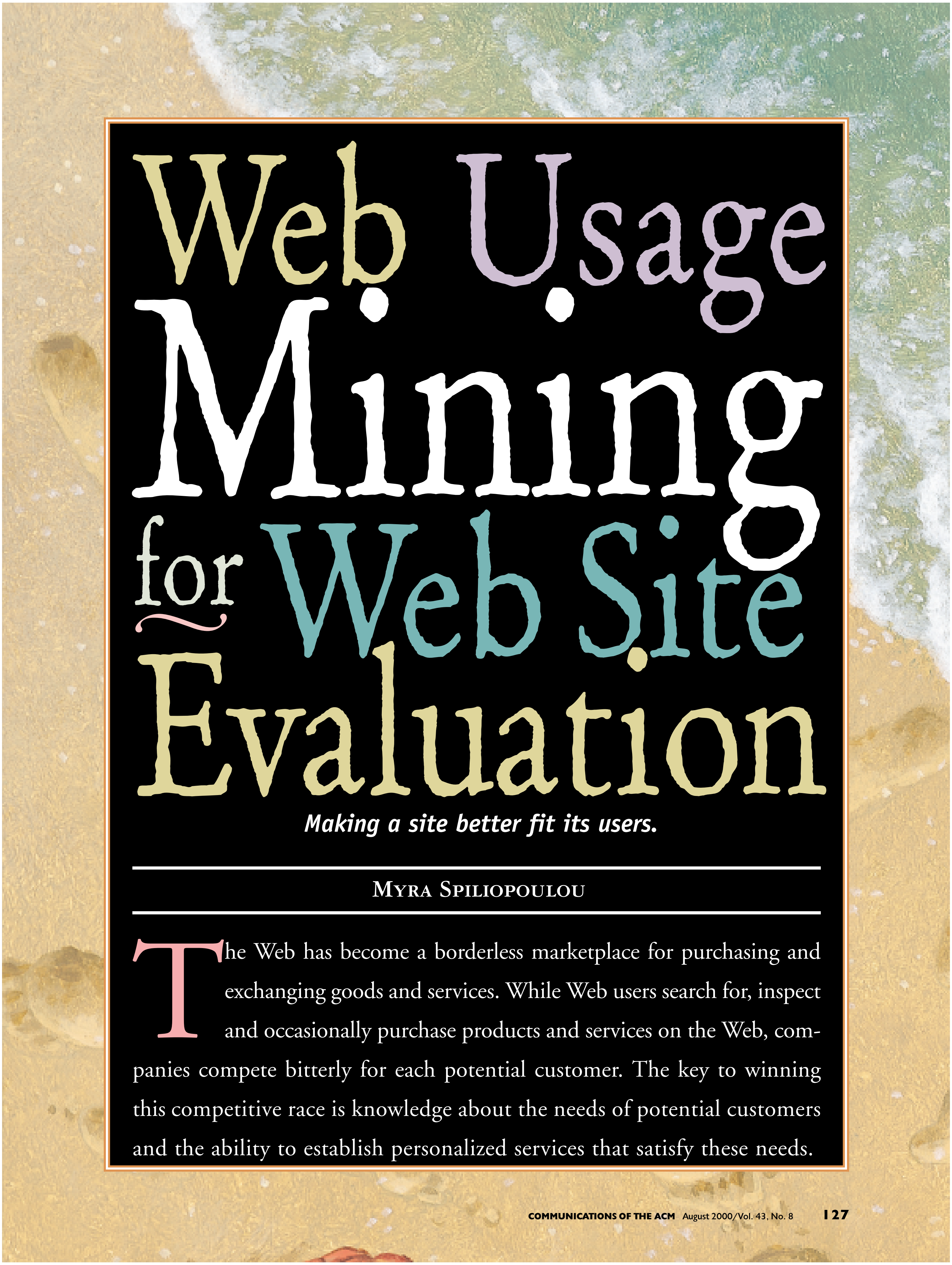
Web Usage Mining for Web Site Evaluation

Article *in* Communications of the ACM · August 2000

DOI: 10.1145/345124.345167 · Source: OAI

Myra Spiliopoulou
Otto-von-Guericke-Universität Magdeburg



Web Usage Mining for Web Site Evaluation

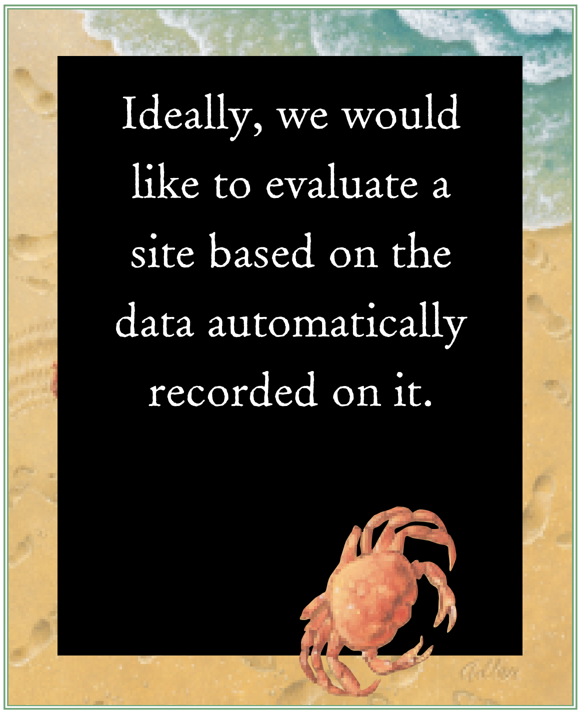
Making a site better fit its users.

Myra Spiliopoulou

The Web has become a borderless marketplace for purchasing and exchanging goods and services. While Web users search for, inspect and occasionally purchase products and services on the Web, companies compete bitterly for each potential customer. The key to winning this competitive race is knowledge about the needs of potential customers and the ability to establish personalized services that satisfy these needs.

The only information left behind by many users visiting a Web site is the trace through the pages they have accessed. From this data source, the site owner must figure out what the users wanted in the site, what they liked and what disturbed or distracted them. It is tempting to conclude that products rarely purchased are of less interest to the potential customers and that the chance of accessing a page is increased by placing a link to it at a prominent place. However, such conclusions are only valid if the users perceive the site and understand its services *as the designers have conceived them.* This is not always the case. Many people are familiar with the procedure of making purchases in a store or of acquiring a document from an authority. This does not imply that adding or removing products from an electronic cart is intuitive to them, or that they can effectively formulate queries to the Web site of a large governmental organization.

Hence, before personalizing the products offered in a Web site to fit the needs of the users, we should personalize the site in serving its users. Otherwise, two evils may occur. First, users having difficulties in understanding how the site should be explored are disappointed—potential customers are lost. Second, their traces blur the statistics about which pages or products are popular or correlated. Invalid conclusions and more confused users could be the result.



Ideally, we would like to evaluate a site based on the data automatically recorded on it.

There are three factors affecting the way a user perceives and valuates a site: content, Web page design, and overall site design. The first factor concerns the goods, services, or data offered by the site. The other factors concern the way in which the site makes content accessible and understandable to its users. We distinguish between the design of individual pages and the overall site design, because a site is not simply a collection of pages—it is a network of related pages. The users will not engage in exploring it unless they find its structure intuitive.

Evaluating a Web Site

Before improving and personalizing a Web site, we need a way of evaluating its current usage. According to Preece et al., "evaluation is concerned with gathering data about the usability of a design or product by a specified group of users for a particular activity within a specified environment or work context" [10].

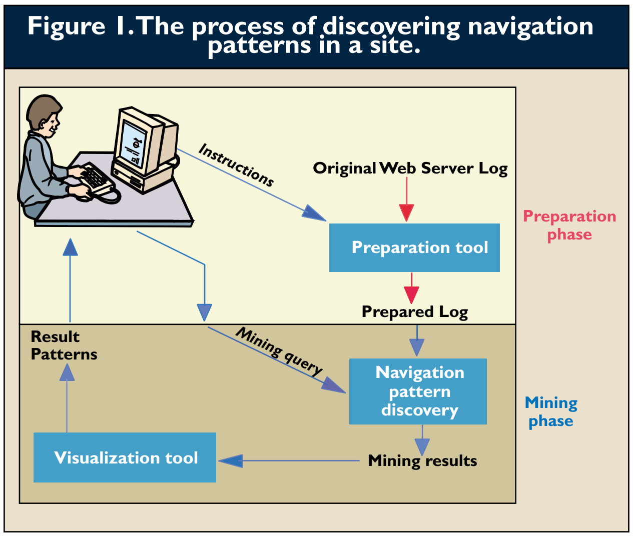
To turn this definition to practice, we first select a group of persons that represent the target group of users addressed by the site. We then define the activities they are expected to perform, such as surfing around, exploring the contents, and ordering products. We record the operations they perform when pursuing these activities. We then can inspect the results informally or analyze them with statistical methods. Further, informal comments from the test persons may be gathered and used in the evaluation.

This methodology is pursued by Eighmey in his field study of multiple commercial Web sites [7]. He established an experimental framework, in which the test persons were confronted with alternative designs of each site. From the three aspects of a site, he considered content and page design. His study concluded that entertainment when accessing a site plays the most important role. The information offered by the site is the second most essential factor. A less positive fact was that none of the investigated sites "succeeded in creating a context and sense of community needed to build a continuing relationship with Web site users."

Thus, in principle, we have an evaluation methodology. But it comes with a high price. We must gather a group of test persons, ensure they are representative of the target group of customers and establish an experimental environment for them to work. Although some companies can afford the time and cost of this process when they establish a site or perform a fundamental redesign, very few can launch the same process each time they want to monitor the site's quality.

Ideally, we would like to evaluate a site based on the data automatically recorded on it. Each site is electronically administered by a Web server, which logs all activities that take place in it in a file, the Web server log. All traces left by the Web users are stored in this log. From this log, we can extract information that indirectly reflects the site's quality by applying data mining techniques.

Figure 1. The process of discovering navigation patterns in a site.



Data Mining for Site Evaluation: Formulating the Problem

Data mining is a methodology for the extraction of knowledge from data. This knowledge is not arbitrary; it relates to a problem, the problem we want to solve. We can perform data mining to optimize the performance of a Web server, to discover which products are being purchased together, or to identify whether the site is being used as expected. The concrete specification of the problem guides us through different preparation and analysis steps of the same Web server log.

The problem of site evaluation leads us to the next questions: What data is appropriate for the analysis? How do we express a concept like "expected usage" to the miner, so that it discovers expected and unexpected navigation patterns? The first question is dealt with in the data preparation phase; the second one is resolved in the mining phase (see Figure 1).

Preparing the Web Log for Analysis

Data mining needs input datasets of high quality in order to produce reliable results. Unfortunately, the data in the original Web server log is of rather low quality and needs extensive preprocessing. For site evaluation, we need a thorough record of the activities of each user in the site, namely all pages visited and the scripts invoked. The first source of difficulty is the absence of user identification data. The user identity is in itself not needed for our analysis. However, we need a means of distinguishing between different users. In market basket analysis, we assume a supermarket cart contains the purchases of a single customer; what would be a supermarket cart for the visitor of a Web site?

In the widely used Common Log Format, the only data recorded about a user is the host name and the version of the user's Web browser. Since many users may access a site from the same host or proxy, we cannot attribute all accesses from this host to the "cart" of a single user. Extended log formats that include the referer URL and the name of the client's agent provide some additional information. However, safe distinction between users can only be guaranteed by means of cookies, scripts, and authentication mechanisms. These mechanisms are not universally accepted by the user community, however. Privacy concerns and, occasionally, security considerations, make some users avoid sites that install cookies or execute scripts on their behalf. Hence, the distinction among users is mostly based on heuristics, a collection of which can be found in [6].

The next source of difficulty is caching. Web clients cache previously visited pages to reduce network traffic and cost. This implies that revisits of the same page are not recorded by the server. However, revisits are an essential characteristic of navigation behavior. Revisits often indicate which pages are observed as related and should become directly connected. Cooley et al. exploit the knowledge on the site's organization to resolve the caching problem [6]: If two pages not directly connected to each other are visited in sequence, then a previously visited page connecting them has been accessed again. If no such page exists, then the accesses come from two different users.

The nature of the site evaluation problem brings up the next issue: Are all recorded activities appropriate for studying how users perceive the site? Perception is a human characteristic. An automated spider browsing through a site does not comprehend it and therefore leaves a spurious navigation pattern. If such visitors access the site frequently, they can statistically blur the human traces.

Martijn Koster has recently made available a database of currently known robots,[1] which can be used for filtering the Web server log. User-programmable agents may be recognized through their non-human behavior, such as depth-first traversal of a whole directory or repeated accesses to the same page every 30 seconds. In the SurfAID project of IBM,[2] the traces of robots are identified by their deviation from typical human-user patterns. Of course, this approach is only safe if we can be sure that no patterns of human users are filtered out for not being typical.

Finally, the prepared log should be combined with further existing information on the site, such as the description of the page contents. Depending on the application, the pages may contain products for sale, advertisements, data extracted from a database, and so forth. Since we are interested in the navigational behavior of the users, the exact objects may be of less impor-

[1] See info.webcrawler.com/mak/projects/robots/active.html

[2] See surfaid.dfw.ibm.com.

tance than the way to find them. Usually, purchasing a red t-shirt requires filling the same forms as for a yellow one, while the forms filled before ordering a bicycle are quite different. Hence, a reasonable step at this point is the abstraction of the objects contained in the pages into "concept hierarchies."

A *concept hierarchy*, also known as taxonomy, generalizes concrete objects into more abstract concepts. Biologists categorize animals according to their anatomy, geographical distribution, habits, and so forth. Librarians use thesauri, which consist of multiple taxonomies, to assign keywords to textual corpora. Concept hierarchies are also useful in data mining, especially for market basket analysis [3]: The analyst groups individual products into more general concepts, with the effect of also grouping purchases of the products together. Thus, associations that are too rare among individual products become apparent when the product groups are studied.

Büchner and Mulvenna suggest concept hierarchies on the users' hosts in order to obtain some demographics on the visitors, such as on the country of origin. They also consider the exploitation of concept hierarchies already available in the enterprise, such as product taxonomies, and propose their incorporation into a Web data warehouse [5]. For the domain of electronic retailing, the SurfAID project of IBM attempts the automatic construction of concept hierarchies: Previously discovered associations among products are combined with text mining techniques that analyze the contents of the Web pages to formulate concepts that describe correlated products.

For the evaluation of form-based sites, such as online catalogs, Berendt and Spiliopoulou propose the establishment of "service-based" concept hierarchies [2]: Instead of generalizing the information contained in the pages into abstract concepts, service-based concept hierarchies model the (combinations of) parametrical settings, with which the services are invoked to generate the pages.

For example, consider a Web site offering access to a collection of scientific articles. The site supports a form-based query service, with which the user can retrieve articles by specifying one or more parameters describing them (title, authors, publisher, journal or conference, keywords, year of appearance, and so forth). In a conventional concept hierarchy, we would abstract the articles being retrieved according to their contents. In a service-based concept hierarchy, we would rather define concepts reflecting the permissible combinations of search parameters like "titleAND-author" or "publisherORjournal." These concepts can be further abstracted in different ways. Thus, concepts like "titleANDauthor," "publisherANDyear," "author-ANDyear" can be abstracted into the more general concept "ConjunctionOfTwoParameters," while a concept like "publisherORjournal" generalizes into a "DisjunctionOfTwoParameters." These concepts can be further generalized into the term "Two-ParametersSearch." This type of hierarchy is appropriate to evaluate the usage of the query service, and in particular to investigate how parameter combinations are being used by the visitors.

Sequences and Sequence Mining

A *sequence* is an ordered list of items, in our case Web pages, ordered by time of access. In the pioneering work of Agrawal and Srikant [1], sequence mining is defined as follows:

> "Given is a collection of transactions ordered in time, where each transaction contains a set of items. The goal is to discover sequences of maximal length that appear more frequently than a given percentage threshold over the whole collection."

A frequent sequence is "maximal," if no sequence containing it is also frequent. If we instruct the miner to find only maximal frequent sequences, we obtain fewer and more compact results.

The definition of the sequence mining problem has an implication: The items constituting a frequent sequence did not necessarily occur adjacently. They just appear in many data records in the same order. This is often desirable: When we investigate the causes of manufacturing errors, we only want the sequences containing error and cause, not the many events in between. The same is true when we search for operating system signals. We will see though, that this implication is impeding when we study the behavior of Web users. **C**

Navigation Pattern Discovery

The data preparation phase restores the users' activities into sequences of page or script accesses. From them, a miner should test whether the site is being used in accordance with the design objectives. However, this informal task is more than what most miners can currently understand.

Sequence miners. The discovery of typical usage patterns seems to be exactly what sequence miners [1, 8] are built for. They discover events (here: accesses to pages) that occur frequently together in the same order. Hence, a sequence miner can find all sequences of Web pages that have been frequently accessed together—see the sidebar “Sequences and Sequence Mining.” However, most of these sequences would be of a trivial nature. For example, imagine a fictitious retailer site, in which the following sequence is the most frequent one:

`Welcome.html –> orders.html` Frequency: 15%

For a retailer site, a person would indeed expect that many users access the Welcome page and later place an order for some products, but the miner cannot have this knowledge. In general, only the designer of the site can say what is trivial and what is not. Thus, the designer is forced to read all patterns discovered by the miner and discard unimportant ones.

It would be much more efficient to automatically test the miner’s results against the expectations of the designer. However, we can hardly expect a site designer to write down all combinations of Web pages that are considered typical; expectations are formed in the human mind in much more abstract terms. Thus, we need miners that can be instructed to do more than just find frequent sequences.

Requirement 1. The miner should understand abstract pattern descriptions, so that the designer can instruct the software on what should be discovered and what should be ignored.

Our rather trivial example sequence contains useful hints. It says that the most frequent route leading to `orders.html` starts at the `Welcome` page. It also says that 85% of the visitors follow other routes. However, this information is not adequate for evaluating and improving the site. The designer would rather prefer answers to questions like: Do most visitors of `orders.html` start at the `Welcome` page or are there many arbitrary routes to it? Can it be that they give up their navigation at some page, and which? To answer those questions, the designer not only needs this sequence, but also the most frequented routes *between the sequence’s ends*.

Requirement 2. A Web usage pattern should be more than a sequence of frequently accessed pages. To reflect the users’ behavior, it should also contain statistics about the routes connecting pages frequently accessed together.

Thus, data mining for site evaluation requires a sophisticated interaction between the site’s designer and the mining software and also a new type of mining software. According to Requirement 2, the expert needs information on frequent patterns; routes adjoining the pages composing the frequent patterns; and statistics on the usage of those routes, so that preferred and rarely used ones can be distinguished.

Web log miners. Dedicated Web log miners form the core of the M*i*DAS mining environment [4] and of the Web Utilization Miner WUM [2, 11]. Both systems have been designed in accordance with the increased demand for intensive interaction with human users, as depicted in Figure 1. This interaction is based on a powerful mining language in which expert users can express their background knowledge, guide the miner and gradually refine or refocus the discovery process, according to the mining results obtained after each query. The mining languages of M*i*DAS and WUM use “templates” to describe many desirable characteristics of the navigation patterns to be discovered, beyond the classic frequency threshold (see the sidebar “Mining Queries in M*i*DAS”). Such characteristics may include a minimum and/or maximum length or Web pages that should or should not appear in the pattern. Thus, Requirement 1 is fullfilled.

A major difference between M*i*DAS and WUM concerns the notion of navigation pattern. In M*i*DAS, a navigation pattern is a sequence of events satisfying the constraints posed by the expert. This is similar to the definition of a frequent sequence in [1, 8]. This implies that only sequences consisting of events frequently occurring together will appear in the mining

Mining Queries in M*i*DAS

M*i*DAS is a Web usage miner designed with the demands of e-commerce applications in mind. It puts particular emphasis on the establishment of a Web data warehouse containing information about the Web site, its contents and its usage. Analysts can specify which patterns are of potential interest to them in a powerful mining language. For example, the analyst of a (fictive) commercial site may formulate the following mining query to express that only those frequent sequences are of interest that (i) start at the `Welcome.html` page and contain a product ordering at the `orders.html` page and (ii) do not contain special offers:

```
<Welcome.html | * | orders.html * >
^ < * | specialOffers.html | * >
```

Without going into the details of the language [4], the first line specifies the two Web pages that must appear in the requested sequences. The second line excludes, through the ^ symbol, sequences containing the `specialOffers.html` page.

The result of this query is a set of frequent sequences, beginning each with an access to `Welcome.html`. Any pages can follow, except for the `specialOffers.html` page, before the other page of interest, `orders.html`, is reached. Each sequence is maximal, in the sense that no sequence contained in it also appears in the result [4].

Navigation Pattern Discovery with WUM

Figure 2. Example Web site.

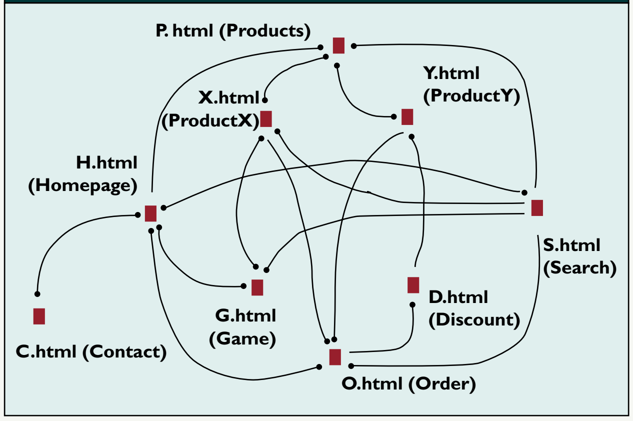


Figure 3. One navigation pattern.

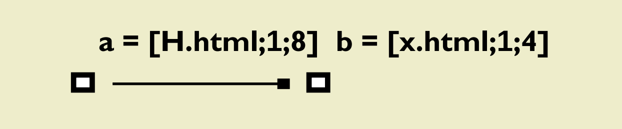
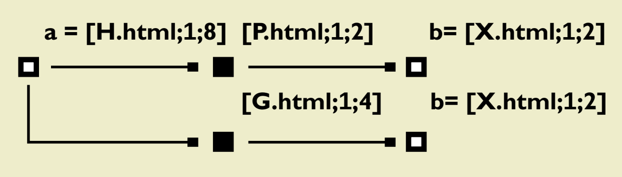


Figure 4. Navigation pattern sequence and tree of roots.



WUM is a Web usage miner designed primarily for the demands of site evaluation, although conventional sequence mining tasks, like correlations of events, can also be performed with it. To demonstrate how the results of WUM give insights into the usage of a Web site, let's consider the fictious site of Figure 2, reproduced from the online demo of WUM (see wum.wiwi.hu-berlin.de). The edges between the pages denote links. We can assume all links to be bidirectional, since the site visitor can always use the Web browser's back button. The diminutive synthetic log consists of 15 sequences. It is not shown here, but can be found at the demo page.

Consider a designer interested in navigation patterns between two pages. The first page should be visited by at least 50% of the users to be of interest. From them, at least 40% should reach the second page. The games page (G.html) may not be the second one, although it may belong to any route between the two pages. This mining query can be expressed in WUM's mining language MINT [2] as follows:

```
SELECT t
FROM NODE AS a b,
TEMPLATE a * b AS t
WHERE a.support > 7
AND (b.support / a.support) ≥ 0.4
AND b.url != "G.html"
```

This query looks like an SQL query. However, the mining results are not records retrieved from a database but groups of sequences constructed by the miner. The letters a, b denote variables, which the miner binds to Web pages. Obviously, the expert does not know in advance which pages will qualify.

The query returns three navigation patterns, one of which is shown in Figure 3. It is comprised of a sequence and of the tree of routes shown in Figure 4. By comparing the structure of the Web site to the sequence of the navigation pattern, we realize that the sequence consists of two nonlinked pages. This is not necessarily a bad sign: linking all products to the home page of the site could be confusing. Rather, users are expected to reach a specific product via a page about all products, P. To verify whether this happens, the designer needs the whole tree of routes between the home page and the product X. There, the designer can see that only two users reach X after visiting P.html. The others go through the games page G.html, an action that may be less expected. C

results. According to Requirement 2 however, the designer also needs information on the routes adjoining the frequent events.

Reflecting this need, the concept of navigation pattern has been extended in WUM to include both the sequence of events that satisfies the expert's constraints *and* the routes connecting those events [11]. The mining result is no more a sequence but a tree composed of these routes. Each page in each route is annotated with the number of visitors that have reached this page via this route. The designer can distinguish between popular and rarely chosen routes by simply inspecting the numbers on the graphical representation of the query result and can also identify pages where users give up their navigation (whenever a popular route mounts to a rarely followed route). Thus, WUM satisfies both Requirement 1 and Requirement 2 on Web usage mining for site evaluation. (See the sidebar "Navigation Pattern Discovery with WUM.")

Evaluating the discovered navigation patterns. Each query to a Web usage miner returns a set of navigation patterns. Then, the analyst faces the nontrivial problem of evaluating these patterns and deriving *reliable* conclusions from them.

According to Requirement 2, a navigation pattern describes one or more routes among given Web pages, along with statistics on how often each page of each route has been accessed. The science of statistics provides rules with which the analyst can determine whether a pattern or a component of it is significant or the output of coincidence. For example, assume a site contains some very popular page P, which users can reach from many other pages. Then, all pages containing links to P will be accessed quite frequently. However, it is not safe to conclude that these pages are also of interest to the users, since their popularity is a reflection of the popularity of P. On the other hand, if only one page leading to P is accessed frequently, while other pages are never used to reach it, this page may itself be of importance to the users.

Statistical testing of the mining results is indispensable. However, site evaluation goes well beyond this test. The reader may recall that the site's designer needs insights regarding which pages should be improved and how. The combination of criteria such as frequency and expectedness of a mining result can help to this end. If the designer sees a frequently followed route and characterizes it as expected, this implies that many users perceives this part of the site as modeled by the designer. If a frequent route is surprising to the designer, this signals that many users navigate differently than originally anticipated when the site was designed. By studying this route closer and comparing it to other routes crossing it, the designer can detect pages that are not properly designed or linked and redesign them.

In [12], we propose a theoretical framework for the evaluation of discovered navigation patterns. We specify the notion of "success" for a Web site as related to the business strategy of its owner: if the site is an online shop, it is successful if the users purchase products; if it is built for product promotion, it is successful if the users click at the advertisements. Then, borrowing from marketing theory, we measure the *conversion efficiency* of a Web page as its contribution to the success of the site: For an online shop, the conversion efficiency of a Web page is the ratio of visitors that purchased a product after visiting this page to the total number of visitors that accessed the page. For a promotional site, the conversion efficiency of the page could be measured as the ratio of visitors that clicked on an advertisement after visiting the page. With this measure of "success" as a basis, the analyst can concentrate on discovered patterns containing pages with low conversion efficiency. These pages should be redesigned to better serve the purposes of the site.

Restructuring a site according to the mining results. Ultimately, navigation pattern discovery should help the designer in improving the site. Restructuring a site by inserting links and redesigning pages does not have solely positive effects, though, since each mining result reflects the navigation behavior of *some* users only and rebuilding the site for them might confuse other users. It is more appropriate to dynamically adapt a page according to the pages visited by the user so far and the pages that the designer recommends as follow-ups. Building recommendations according to mining results is discussed by Mobasher et al. in their article in this special section.

Site Evaluation in Practice

Web usage analysis is a long process of learning to see a Web site from the perspective of its users. We applied the miner WUM on several Web sites. Our preliminary results on one of them are presented in [2, 12] and summarized here.

The SchulWeb site[3] accommodates the largest and most comprehensive database of German secondary education schools on the Web, a database of German language high school magazines on the Web, and a collection of online resources, and communication services. Access to the site's resources occurs via form-based queries. The query results are records retrieved from the database and placed in dynamically generated pages.

For the analysis of this site, Berendt and Spiliopoulou have concentrated on the group of users looking for a particular school. The corresponding query interface allows the specification of (a) the region, where the school should be located, (b) the school type, currently choosing among three types and (c) a text string that should be contained into an attribute of the school, for example, the school name (default!), the name of a teacher in this school, the town and so forth. The concept hierarchies built for the analysis reflected the possible combinations of search parameters. In particular, the search options were abstracted into "RegionalSearch," "SchoolTypeSearch" and "TextSearch," and concepts were devised for the permissible parameter combinations.

According to the site's design, three steps suffice to reach an individual school. Occasionally, paging over a

[3] See "School Web" at www.schulweb.de.

long list of schools can be expected. Hence, it was tested whether users reach a school in a *short* number of steps. In particular, WUM was used to discover the following types of frequent navigation patterns that reflect the *conversion efficiency* of the search strategies offered by the SchulWeb.

First, frequent navigation patterns were discovered that involved a search strategy and the subsequent few steps until a page was reached with relatively high confidence [2]. This page was not always a school, since many users needed much more than three steps to reach a school after starting their search. However, the activities performed before reaching a school were of interest: they reflected the attempts of the users to improve the search result by changing or refining the search strategy. Second, WUM was guided to discover frequent navigation patterns leading to a school within a small number of steps. Then, patterns with similar content but not leading to a school were identified and compared to the former [12].

The discovered navigation patterns showed that users prefer search parameters that allow them to select a value by clicking, instead of typing text themselves. Also, they mostly perform regional searches, that is, they are interested in schools of some particular region. To this purpose, they pose an initial, rather unrestrictive query that retrieves all schools in the region. Then they try to reduce the result size by extending the query with additional parameter specifications. This effort does not always produce a result list of manageable size, so that users often browse through long lists of results.

In response to these findings [2], the query interface was modified to better support regional searches. In particular, the default parameter for text search is now the town name, instead of the name of the school. The preliminary results of a new mining session showed that the modified interface leads to a more successful usage of the site [12].

Conclusion

Personalized Web access services are a demand of many users that feel overwhelmed with the information available on the Web. Building a site that satisfies this demand presupposes knowing the site as it is perceived by its users. This knowledge is not trivial to acquire, because the site designer has a different perception of the site's content and intended use than the occasional, the regular, the novice, or the expert user. To personalize a site according to the requirements of each user, user navigation patterns must be discovered and analyzed.

The navigation patterns reflect how the site is being perceived by different groups of users. It is not possible to establish a static site that satisfies all groups. Instead, a service should assist the user by finding expected and unexpected patterns that contain the user's trace thus far and adjust the content of each page in such a way that the expected route becomes apparent and the unexpected route is still possible. Services for adaptive page generation do exist. They must be adapted to this new use.

Web usage analysis extracts knowledge from a Web server log. The research on data preparation and data mining in this domain already contains many remarkable contributions. It is anticipated, though, that Web usage mining is particularly difficult due to a gap between the advances of human-computer interaction, the practice of Web site organization, and the support offered by Web usage miners in evaluating the quality of a site.

To close this gap, mining should better reflect the knowledge of the designer on Web site usage. This knowledge must be injected in the preparation of the data, in the instructions of the analyst to the miner, and in the interpretation of the results. To the latter task, the analyst and the designer are currently assisted by visualization tools and statistic theory. This must be enhanced by a better understanding of the results by the mining software itself.

References

1. Agrawal, R. and Srikant, R. Mining sequential patterns. In *Proceedings of the International Conference on Data Engineering* (Taipei, Taiwan, Mar. 1995).
2. Berendt, B. and Spiliopoulou, M. Analyzing navigation behavior in Web sites integrating multiple information systems. *VLDB Journal*, Special Issue on Databases and the Web 9, 1 (2000), 56–75.
3. Berry, M. and Linoff, G. *Data Mining Techniques for Marketing, Sales, and Customer Support.* Wiley, NY, 1997.
4. Büchner, M., Baumgarten, M. Anand, S.S., Mulvenna, M.D., and Hughes, J.G. Navigation pattern discovery from Internet data. In *WEBKDD'99*, San Diego, CA, (August 1999).
5. Büchner, A.G. and Mulvenna, M.D. Discovering Internet marketing intelligence through online analytical Web usage mining. *ACM SIGMOD Record* (Dec. 1998), 54–61.
6. Cooley, R., Mobasher, B., and Srivastava, J. Data preparation for mining World Wide Web browsing patterns. *Journal of Knowledge and Information Systems 1*, 1 (1999).
7. Eighmey, J. Profiling user responses to commercial Web sites. *Journal of Advertising Research 37*, 2 (May–June 1997), 59–66.
8. Mannila, H. and Toivonen, H. Discovering generalized episodes using minimal occurrences. In *Proceedings of the 2nd International Conference KDD'96*, (1996), 146–151.
9. Masand, B. and Spiliopoulou, M, Eds. KDD'99 Workshop on Web Usage Analysis and User Profiling WEBKDD'99, San Diego, CA, Aug. 1999. ACM. Online archive of the extended abstracts at www.acm.org/sigkdd/proceedings/webkdd99/.
10. Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T. *Human-Computer Interaction.* Addison Wesley, 1994.
11. Spiliopoulou, M. The laborious way from data mining to Web mining. *International Journal of Computer Systems, Science, and Engineering, Special Issue on Semantics of the Web 14* (Mar. 1999), 113–126.
12. Spiliopoulou, M. and Pohle, C. Data mining for measuring and improving the success of Web sites. *Data Mining and Knowledge Discovery, Special Issue on Electronic Commerce,* 2000.

Myra Spiliopoulou (myra@wiwi.hu-berlin.de) is an assistant professor in the Institute of Information Systems at Humboldt University in Berlin; www.wiwi.hu-berlin.de/~myra.