Automatic Personalization Based on Web Usage Mining

Web usage mining can help improve the scalability, accuracy, and flexibility of recommender systems.

BAMSHAD MOBASHER, ROBERT COOLEY, AND JAIDEEP SRIVASTAVA

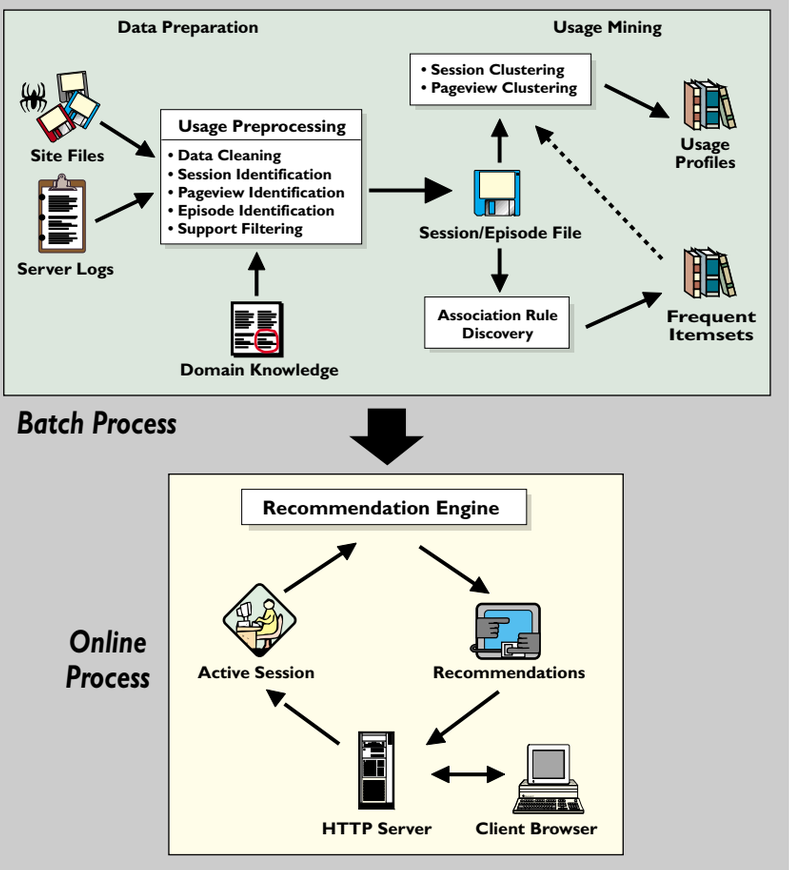
The ease and speed with which business transactions can be carried out over the Web have been a key driving force in the rapid growth of e-commerce. The ability to track user browsing behavior down to individual mouse clicks has brought the vendor and end customer closer than ever before. It is now possible for vendors to personalize their product messages for individual customers on a massive scale, a phenomenon referred to as "mass customization." Of course, this type of

personalization is applicable to any Web browsing activity, not just e-commerce. Web personalization can be defined as any action that tailors the Web experience to a particular user, or set of users. The experience can be something as casual as browsing a Web site or as (economically) significant as trading stocks or purchasing a car. The actions can range from simply making the presentation more pleasing to anticipating the needs of a user and providing customized information.

To date, most personalization systems for the Web have fallen into three major categories: manual decision rule systems, collaborative filtering systems, and content-based filtering agents. Manual decision rule systems, such as Broadvision (www.broadvision.com), allow Web site administrators to specify rules based on user demographics or static profiles (collected through a registration process), or session history. The rules are used to affect the content served to a particular user. Collaborative filtering systems, such as Firefly [11], and Net Perceptions (www.netperceptions.com), typically take explicit information in the form of user ratings or preferences, and, through a correlation engine, return information that is predicted to closely match the users' preferences. Content-based filtering approaches such as those used by WebWatcher [5] rely on content similarity of Web documents to personal profiles obtained explicitly or implicitly from users.

Increasingly, the new generation of Web personalization tools is attempting to incorporate techniques for pattern discovery from Web usage data. For example, some collaborative filtering systems such as Net Perceptions are experimenting with obtaining implicit user ratings from usage data. Web usage mining systems run any number of data mining algorithms on usage or clickstream data gathered from one or more Web sites in order to discover user profiles. The increasing focus on Web usage data is due to several factors. The input is not a subjective description of the users by the users themselves, and thus is not prone to biases. The profiles are dynamically obtained from user patterns, and thus the system performance does not degrade over time as the profiles age. Furthermore, using content similarity alone as a way to obtain aggregate profiles may result in missing important semantic relationships among Web objects. Thus, Web usage mining can reduce the need for obtaining subjective user ratings or registration-based personal preferences.

Figure 1. A general architecture for usage-based Web personalization.



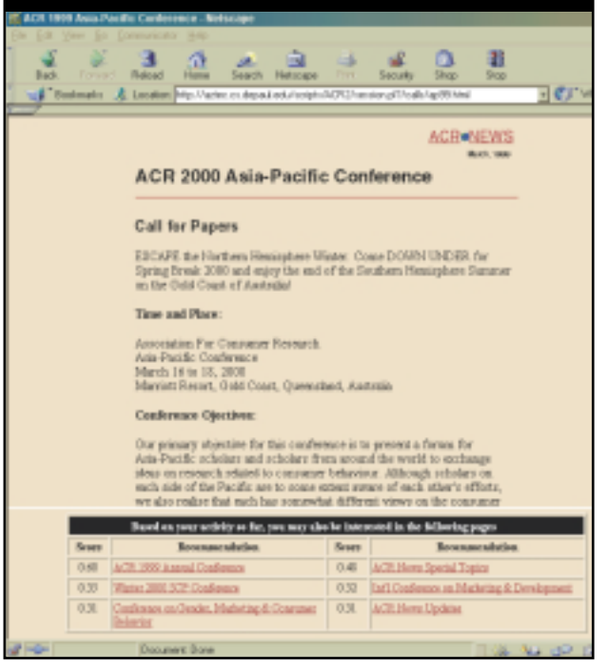
Mining Usage Data for Web Personalization

Principal elements of Web personalization include the modeling of Web objects (for example, products or pages) and subjects (users), categorization

Figure 2. Main page for the demonstration site. Initially, no recommendations are provided as the active user session does not contain sufficient number of references.

Figure 3. Dynamic recommendations after the user has navigated through "President's Column" and "Online Archives" pages.

Figure 4. The system provides specific recommendations related to conferences based on user navigation through "Conference Update," "Call for Papers," and "Asia Pacific Conference" pages.



of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization. As depicted in Figure 1, the overall process of usage-based Web personalization is divided into two components. The offline component is comprised of the data preparation and specific usage mining tasks. The data preparation tasks result in a server session file, where each session is a sequence of pageviews each represented by a unique Uniform Resource Identifier (URI) reference attributed to a particular user. In addition, only URIs that represent meaningful or relevant pageviews are included in a server session file (see the sidebar "Data Preparation for Web Usage Mining"). The usage mining tasks can involve the discovery of association rules, sequential patterns, pageview clusters, user clusters, or any other pattern discovery method. The discovered patterns are used by the online component to provide personalized content to users based on their current navigational activity. The personalized content can take the form of recommended links or products, targeted advertisements, or text and graphics tailored to the user's preferences. The Web server keeps track of the active server session as the user's browser makes HTTP requests. The recommendation engine considers the active server session in conjunction with the discovered patterns to provide personalized content.

Data preparation. The prerequisite step to any type of usage mining is the identification of a set of server sessions from the raw usage data. Ideally, each server session gives an exact accounting of who accessed the Web site, what pages were requested and in what order, and how long each page was viewed. Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into various data abstractions (see the sidebar "Data Preparation for Web Usage Mining"). The practical difficulties in

Table 1. User behavior profiles.

	Weight	Pageview URI
Profile 1	0.78	Call for Papers
	0.67	CFP: ACR 1999 Asia-Pacific Conference
	0.64	CFP: Society For Consumer Psychology Conference
	0.61	ACR 1999 Annual Conference
	0.55	CFP: ACR 1999 European Conference
	0.52	CFP: Int'l Conference on Marketing and Development
	0.50	Conference Update
Profile 2	0.82	CFP: Journal of Psychology and Marketing II
	0.71	CFP: Society For Consumer Psychology Conference
	0.68	Conference Update
	0.68	CFP: Journal of Consumer Psychology II
	0.56	CFP: Conference on Gender, Marketing and Consumer Behavior
	0.52	Online Archives

Table 2. Commonly used data abstractions for Web usage mining.

Term	Definition
User	Single individual that is accessing files from one or more Web servers through a browser.
Page File	File that is served through HTTP protocol to a user.
Pageview	Set of page files that contribute to a single display in a Web browser.
Server Session	Set of pageviews served due to a series of HTTP Requests from a single user to a single Web server.
Episode	Subset of pageviews from a single server session.

performing preprocessing are a moving target. As the technology used to deliver content over the Web changes, so do the preprocessing challenges. While each of the basic preprocessing steps remains constant, the difficulty in completing certain steps has changed dramatically as Web sites have moved from static HTML served directly by a Web server to dynamic scripts created from sophisticated content servers and personalization tools. Both client-side tools (browsers) and server-side tools (content servers) have undergone several generations of improvements since the inception of the Web.

Discovery of usage profiles. The session file obtained in the data preparation stage can be used as the input to a variety of data mining algorithms such as the discovery of association rules or sequential patterns, clustering, and classification. At this point in the process, the results of the pattern discovery can be tailored toward several different aspects of Web usage mining. For example, Perkowitz and Etzioni [8] have proposed the idea of dynamically creating multiple index pages for a site based on co-occurrence patterns of pages among user sessions. Schechter et al. [10] have developed techniques for using the path profiles of users to predict future HTTP requests, which can be used for network and proxy caching. Spiliopoulou et al. [9], Cooley et al. [2], and Buchner and Mulvenna [1] have applied data mining techniques to extract usage patterns from Web logs for the purpose of deriving marketing intelligence. Shahabi et. al [12] and Nasraoui et al. [6] have proposed clustering of user sessions to predict future user behavior.

However, the discovery of patterns from usage data by itself is not sufficient for performing the personalization tasks. The critical step is the effective derivation of good quality and useful (that is, actionable) "aggregate profiles" from these patterns. Ideally, profiles capture aggregate views of the behavior of subsets of users based on their interests and/or information needs. In particular, aggregate profiles must exhibit three important characteristics—they should:

- Capture possibly overlapping interests of users, since many users may have common interests up to a point (in their navigational history) beyond which their interests diverge;
- Provide the capability to distinguish among pageviews in terms of their significance within the profile; and
- Have a uniform representation that allows for the recommendation engine to easily integrate different kinds of profiles (multiple profiles based on different pageview types, or obtained via different mining techniques).

Given these requirements, we have found that representing usage profiles as weighted collections of URIs provides a great deal of flexibility. Each item in a usage profile is a URI uniquely representing a relevant pageview, and can have an associated weight representing its significance within the profile. The usage profiles can be viewed as ordered collections (if the goal is to capture the navigational path profiles followed by users [9]), or as unordered collections (if the focus is on capturing associations among specified content or product pages). Based on the information collected for each pageview during preprocessing, other types of constraints can also be imposed on profiles (for example, we may wish to focus the personalization effort only on certain types of products or pages related to specific content cate-

gories). Another advantage of this representation is that the profiles themselves can be viewed as vectors, thus facilitating the task of matching a current user session with similar profiles using standard vector operations.

Traditional collaborative filtering techniques are often based on real-time matching of the current user's profile against similar records (nearest neighbors) obtained by the system over time from other users. However, as noted in recent studies [7], it becomes hard to scale collaborative filtering techniques to a large number of items (for example, pages or products), while maintaining reasonable prediction performance and accuracy. Part of this is due to the increasing sparsity in the data as the number of items increase. One potential solution to this problem is to first cluster user records with similar characteristics, and focus the search for nearest neighbors only in the matching clusters. In the context of Web personalization, this task involves clustering user sessions identified in the preprocessing stage.

A variety of clustering techniques can be used for clustering similar sessions based on occurrence patterns of URI references. User sessions can be mapped into a multidimensional space as vectors of URI references (so, the dimensions—or features—are the URIs appearing in the session file). Standard clustering algorithms generally partition this space into groups of items that are close to each other based on a measure of distance or similarity. Dimensionality reduction techniques may be employed to focus only on relevant or significant features. For example, support filtering discussed earlier (see the sidebar "Data Preparation for Web Usage Mining") can provide an effective dimensionality reduction

Data Preparation for Web Usage Mining

When dealing with server-side data collection from a Web server log or packet-sniffer, the major difficulties in usage preprocessing are due to the incompleteness of the available data. The required high-level tasks are data cleaning, user identification, session identification, pageview identification, and path completion. In addition, episode identification can be optionally performed as a final preprocessing step prior to pattern discovery. Some amount of content and structure preprocessing is almost always necessary. A list of commonly used terms associated with preprocessing is shown in Table 2; the figure appearing here summarizes the preprocessing steps.

Data cleaning is usually site-specific, and involves tasks such as merging logs from multiple servers, removing graphics file accesses, and parsing of the logs. The difficulties involved in identifying users and sessions depends greatly on the server-side technologies used for the Web site. For Web sites using cookies or embedded session IDs, user and session identification is trivial. Web sites without the benefit of additional information for user and session identification must rely on heuristics, such as those presented in [2].

Pageview identification is the task of determining which page file accesses contribute to a single pageview, and is heavily dependent on the intrapage structure. For a single frame site, each HTML file has a one-to-one correlation with a pageview. However, for multiframed sites, several files make up a given pageview. Without detailed site structure information, it is very difficult, if not impossible, to infer pageviews from a Web server log. Not all pageviews are relevant for specific mining tasks, and among the relevant pageviews some may be more significant than others. The significance of a pageview may depend on usage, content and structural characteristics of the site, as well as on prior domain knowledge specified by the site designer and the data analyst. For example, in an e-commerce site pageviews corresponding to product-oriented events (for example, shopping cart changes or product information views) may be considered more significant than others. Similarly, in a site designed to provide content, content pages may be weighted higher than navigational pages. In order to provide a flexible framework for a variety of data mining activities a number of attributes must be recorded with each pageview. These attributes include the pageview id (normally a URI uniquely representing the pageview), duration, static pageview type (information page, product view, index page, and so forth), and other metadata, such as content attributes.

Path completion involves inferring cached pageviews based on the referring information from the logs. The extent of the problem created by caching is dependent on both the server-side and client-side technologies. Dynamic content with unique URIs for each server session (An embedded session ID is a common method for making URIs unique) will not be subject to proxy level caching. However, any type of content can be cached at the client level. The amount of client level caching is set by the client-side browser.

It is also possible to obtain a further level of granularity by identifying *episodes* within the sessions [2] (episodes are referred to as transactions in [2]). The goal of episode identification is to dynamically create

method while actually improving clustering results. Ideally, each cluster represents a group of users with similar navigational patterns. However, session clusters by themselves are not an effective means of capturing an aggregated view of common user profiles. Each session cluster may potentially contain thousands of user sessions involving hundreds of URI references. In our Web usage mining framework, the ultimate goal in clustering user sessions is to obtain actionable usage profiles which, as noted previously, can be represented as weighted collections of URIs. We discuss one method for obtaining useful profiles from session clusters in the discussion of the WebPersonalizer.

The representation of user sessions as vectors of URI references can provide a number of advantages and a great deal of flexibility. For instance, the distance or similarity among sessions can be computed using standard vector operations. Furthermore, depending on the goals of Web usage mining, a variety of weights can be chosen for each URI in a session vector. Weights can be based on the amount of time users spend on pages referenced by each URI, or they can be based on prior domain knowledge specified by the site owner (for example, in an online catalog, the site owner may wish to weigh product pages referenced by URIs more heavily than other informational pages within the site).

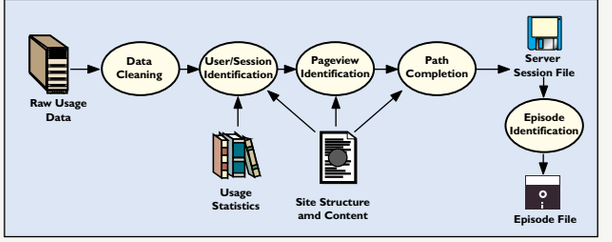
For example, consider the two usage profiles derived from session clusters of the site for Association for Consumer Research shown in Table 1 (also see the sidebar "Experiments with the WebPersonalizer System"). In Table 1, Profile 1 captures the behavior of users interested in current and upcoming conferences during 1999 related to consumer research. On the other hand, Profile 2 captures the

meaningful clusters of references for each user, based on an underlying model of the user's browsing behavior. This allows each page reference to be categorized as a *content* or *navigational* reference for a particular user. Content references can be further classified according to page types or the type of user activity (for example, product purchases). For the purpose of this article, however, we focus on server sessions as the units of user activity to which various data mining techniques are applied.

Content preprocessing consists of converting the text, image, scripts, and other multimedia files into forms that are useful for the Web Usage Mining process. Often, this consists of performing content mining such as classification or clustering. While applying data mining to the content of Web sites is an interesting area of research in its own right, in the context of Web Usage Mining, the content of a site can be used to filter the input to, or output from the pattern discovery algorithms. For example, results of a classification algorithm could be used to limit the discovered patterns to those containing pageviews about a certain subject or class of products. In addition to classifying or clustering page views based on topics, pageviews can also be classified according to their intended use. Pageviews can be intended to convey information (through text, graphics, or other multimedia), gather information from the user, allow navigation (through a list of hypertext links), or some combination uses. The intended use of a pageview can also filter the sessions before or after pattern discovery.

The structure of a site is created by the hypertext links between pageviews. The structure can be obtained and preprocessed in the same manner as the content of a site. This is necessary for handling pageviews that have multiple frames, dynamic pages that have the same template name for multiple page views, as well as dealing with extraneous references such as to image or sound files. It may also be necessary to filter the log files by mapping the references to

Summary of the preprocessing steps.



the site topology induced by physical links between pages. This is particularly important for usage-based personalization, since the recommendation engine should not provide dynamic links to "out-of-date" or non-existent pages.

Finally, the session file can be filtered by removing very small server sessions or episodes, and low-support URI references—references to those URIs that do not appear in a sufficient number of sessions. This type of support filtering can be useful in eliminating noise from the data, such as that generated by shallow navigational patterns of "non-active" users, and URI references with minimal knowledge value for the purpose of personalization. C

behavior of users who are more specifically interested in conferences and journals related to consumer psychology. Note that the behavior of a single user may match both profiles during the same or different sessions.

Another approach for obtaining aggregate usage profiles is to directly compute (overlapping) clusters of pageview references based on how often they occur together across user sessions (rather than clustering sessions, themselves). We call the usage profiles obtained in this way pageview clusters. In general, this technique will result in a different type of aggregate profiles as compared to the session clustering technique. The usage profiles derived from session clusters group together pages that co-occur commonly across similar sessions. On the other hand, pageview clusters tend to group together frequently co-occurring items across sessions, even if these sessions are themselves not deemed to be similar. This technique allows one to obtain clusters that potentially capture overlapping interests of *different* types of users. The question of which type of clusters are most appropriate for personalization tasks is an open research issue. However, the answer to this question, in part, depends on the structure and content of the specific site, as well as the goals of personalization actions.

The difficulty in clustering URIs directly comes from the high dimensionality of the feature space. The user sessions, measured in tens to hundreds of thousands in a typical application, must be used instead of the URIs as features. Traditional clustering techniques, such as distance-based methods, generally cannot handle this type of clustering. Furthermore, dimensionality reduction in this context may not be appropriate, as removing a significant number of sessions as features may result in losing too much information. In the next section we discuss an approach based on *Association Rule Hypergraph Partitioning*, which has been found to

Experiments with the WebPersonalizer System

To demonstrate the feasibility of the proposed techniques and architecture, we conducted a series of experiments using the site for the newsletter of the Association for Consumer Research (acrnews.org). The site contains a variety of news items, including President's columns, conference announcements, and calls-for-papers for a number of conferences and journals. The site was used to implement a demonstration version of the WebPersonalizer system based on the architecture presented here. A local version of the demonstration site is available from aztec.cs.depaul.edu/scripts/ACR2.

We used a subset of the ACR logs (from June 1998 to June 1999), and used session clustering to derive usage profiles. The session file for this experiment contained 18,430 user sessions with a total of 192 unique URIs. Based on the average session size, a session window of size 3 was chosen. The session clustering process yielded 28 usage profiles representing different types of user access patterns. A threshold of 0.5 was used to derive usage profiles from session clusters (that is, usage profiles contained only those URI references appearing in at least 50% of sessions). We used a recommendation threshold of 0.3 as a cutoff point to ensure capturing overlapping user interests.

The recommendation engine was implemented as a set of CGI scripts, using cookies to keep track of user's active session. The figures appearing here depict a typical interaction of user with site. The top frame in each window contains the actual page contents from the site, while the bottom frame contains the recommended links. When the user clicks on a link in either frame, the top frame will display the content of the requested page, and the bottom frame is dynamically updated to include the new recommendations. As seen in Figure 2, initially the system does not provide any recommendations until the user has navigated through more pages. Figure 3 shows the recommendations resulting after the user has followed a path to "President's Column" and then to "Online Archives." The recommendations include past President Columns and Editor's Notes (as well as other pages) often visited by users who have shown similar access patterns. Figure 4 shows the results of the user navigation through "Conference Update," "Call for Papers," and then "1999 Asia Pacific Conference." As can be seen in these Figures, user's intention of looking for more specific information will result in more specific recommendations. For example, when the user accesses a specific conference page (Figure 4), other specific conference information is presented as potentially interesting (for example, "Winter 2000 SCP Conference" and "Int'l Conference on Marketing and Development").

Additional experimental results with other datasets comparing the various techniques described in this article can be found at maya.cs.depaul.edu/~mobasher/personalization/. C

be particularly suitable for this task. Another approach for the clustering URIs directly may be based on the cluster mining technique of Perkowitz and Etzioni (see their article "Adaptive Web Sites" in this issue).

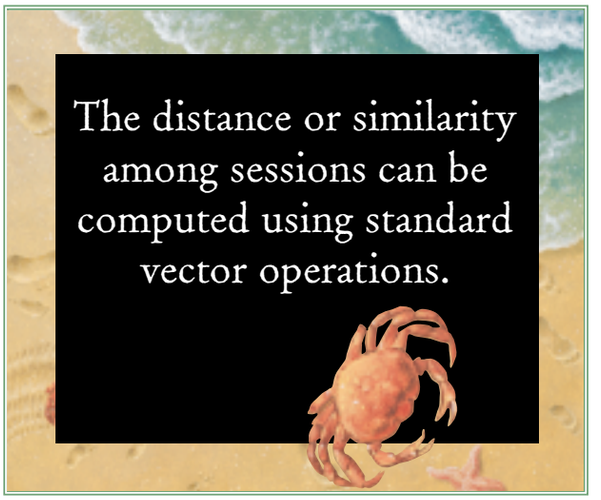
From profiles to recommendations. The recommendation engine is the online component of a Web personalization system. The task of the recommendation engine is to compute a *recommendation set* for the current (active) user session, consisting of the objects (links, ads, text, products, and so forth) that most closely match the current user profile. The essential aspect of computing a recommendation set for a user is matching the current user's activity against aggregate usage profiles. The recommendation engine must be an online process, providing results quickly enough to avoid any perceived delay by the users (beyond what is considered normal for a given Web site and connection speed).

If the data collection procedures in the system include the capability to track users across visits, then the recommendation set can represent a longer term view of potentially useful links based on the user's activity history within the site. On the other hand, if profiles are derived from anonymous user sessions contained in log files, then the recommendations provide a short-term view of user's navigational history. As depicted in Figure 1, these recommended objects are then added to the last page in the active session accessed by the user before that page is sent to the browser.

In general there are several design factors that can be taken into account in determining the recommendation set. These factors may include:

- A short-term history depth for the current user representing the portion of the user's activity history that should be considered relevant for the purpose of making recommendations;
- The mechanism used for matching aggregate usage profiles and the active session; and
- A measure of significance for each recommendation (in addition to its prediction value), which may be based on prior domain knowledge or structural characteristics of the site.

Maintaining a history depth is important because most users navigate several paths leading to independent pieces of information within a session. In many cases these *episodes* have a length of no more than two or three references. In such a situation, it may not be appropriate to use references a user made in a previous episode to make recommendations during the current episode. It is possible to capture the user history depth within a sliding window over the current session.



A variety of techniques can be used to match the active user session with one or more of the discovered usage profiles. For instance, standard classification techniques can be employed to automatically assign the new user session to a class determined based on aggregate profiles. It is also possible to directly use patterns discovered as part of the association rule (or sequential pattern) discovery to provide recommendations (see the sidebar "Mining Association Rules for Personalization"). In the architecture described in this article, the aggregate profiles are represented as weighted URI collections. This will allow for both the active session and the profiles to be treated as n-dimensional URI vectors, where n is the number of URI references appearing in the session file. In this case, standard measures of distance or similarity can be utilized to match the active session and the usage profiles, and the recommendations can be ranked according to a matching score. This is the method we have used in the WebPersonalizer system.

Finally, structural characteristics of the site or prior domain knowledge can be used to associate an additional measure of significance with each recommendation. For instance, the site owner or the site designer may wish to consider certain page types (content versus navigational) or product categories as having more significance in terms of their recommendation value. In this case, significance weights can be specified as part of the domain knowledge. Or, it may be desirable to consider pages that are farther away from the current user location within the site as being better recommendations. In this case, structural information such as the link distances can be used to provide significance weighting for recommendations.

The WebPersonalizer System

The WebPersonalizer system uses the architecture shown in Figure 1 to provide a list of recommended hypertext links to a user while browsing through a Web site. Currently, the WebPersonalizer system relies solely on anonymous usage data provided by Web server logs and the hypertext structure of a site. The preprocessing steps outlined in [2] are used to convert the server logs into server sessions. Two different methods, each with its own characteristics, are used to discover aggregate usage profiles represented by a set of URIs. The first method involves the computation of session clusters and the derivation of useful aggregate user profiles from these session clusters. In the second method, we use frequent itemsets discovered as part of association rule discovery to directly obtain clusters of URIs based on their usage characteristics (pageview clusters). Once the representative usage profiles have been computed, a partial session for the current user (the active session) can be assigned to one or more matching usage profiles. The matching profiles are used as the basis for providing the user with additional recommendations.

In order to derive usage profiles from each session cluster, the cluster centroids (the mean vectors) are computed. The mean value for each URI in the mean vector is computed by finding the ratio of the number of occurrences of that URI across all sessions to the total number of sessions in the cluster. Then, the low-support URIs (those with mean value below a certain threshold), are filtered out. For example, if the threshold is set at 0.5, then each usage profile will contain only those URI references that appear in at least 50% of the sessions within its associated session cluster.

For the second method (computing usage profiles directly), the WebPersonalizer system uses the Association Rule Hypergraph Partitioning (ARHP) technique [4]. ARHP is well-suited for this task since it can efficiently cluster high-dimensional data sets without requiring dimensionality reduction as a preprocessing step. Furthermore, the ARHP provides automatic filtering capabilities, and does not require distance computations. The ARHP has been used successfully in a variety of domains, including the categorization of Web documents [3]. In this method the set of frequent itemsets are used as hyperedges to form a hypergraph. A hypergraph is an extension of a graph in the sense that each hyperedge can connect more than two vertices. The weights associated with each hyperedge are computed based on the confidence of the association rules involving the items in the frequent itemset. The hypergraph is then recursively partitioned into a set of clusters. The similarity among items is cap-

Mining Association Rules for Personalization

Association rules[1] capture the relationships among items based on their patterns of co-occurrence across transactions in transactional databases such as point-of-sale data collected in supermarkets. In the case of Web transactions, association rules capture relationships among URI references based on the navigational patterns of users. For example, an association rule

{A.html, B.html} ->
{C.html} [support = 0.01, confidence = 0.75]

represents the relationship that users who access pages A.html and B.html also tend to (with a confidence of 75%) access the page C.html. The support value represents the fact that the itemset {A.html, B.html, C.html} was present in 1% of user sessions. Association rule discovery methods initially find groups of items (which in this case are the URIs appearing in the preprocessed log) occurring frequently together in many transactions), satisfying a minimum support criteria. Such itemsets are referred to as frequent itemsets.

It is possible to consider the frequent itemsets discovered as part of association rule mining directly as usage profiles. The current user session can be matched against frequent itemsets to find candidate recommendations: if the rule satisfies a specified confidence threshold, then the candidate URI is added to the recommendation set. It is also possible to extend this technique to *sequential patterns*[2], particularly when the focus is capturing user navigational paths (see also "Personalizing a Site with Web Usage Mining" in this issue). The problem with this method is that it might be difficult to find large enough itemsets with sufficient support to match the current session. This is particularly true for sites with very small average session sizes. An alternative to reducing the support threshold in such cases would be to reduce the session window size. This latter choice may itself lead to some undesired effects since we may not be taking enough of the user's activity history into account. C

[1]Agrawal, R. and Srikant, R. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB* conference (Santiago, Chile, 1994), 487–499.
[2]Agrawal, R. and Srikant, R. Mining sequential patterns. In *Proceedings of the International Conference on Data Engineering (ICDE)*, (Taipei, Taiwan, Mar. 1995).

tured implicitly by the frequent item sets. Each cluster represents a group of items (URIs) that are very frequently accessed together across sessions. The *connectivity* value of vertex (a URI appearing in the frequent item set) with respect to a cluster measures the percentage of edges with which a vertex is associated. The significance weight of the URI within the resulting profile is obtained as a function of the connectivity value for that URI.

In the case of usage profiles derived from session clustering, the weight for a URI is its mean value in the cluster mean session vector. In the case of pageview clusters obtained using the ARHP method, the weight is the connectivity value of the item within the cluster. In computing the matching scores, the system normalizes for the size of the clusters and the active session. This corresponds to the intuitive notion that we should see more of the user's active session before obtaining a better match with the larger cluster. Furthermore, a candidate URI is considered to be a better recommendation if it is farther away form the current active session. To capture this notion, the *physical link distance* between the active session and a URI is measured (this is the smallest path in the site graph between the URI and any of the URIs in the session).

The full recommendation set for current active session is computed by collecting all URIs whose recommendation score satisfies a minimum threshold requirement from each matching profile. The URIs in the recommendation set are ranked according to their recommendation score when presented to the user. Details of the specific techniques used in the recommendation process, as well as a set of experiments comparing them can be found at maya.cs.depaul.edu/~mobasher/personalization/.

Conclusion

The Web is providing a direct communication medium between the vendors of products and services, and their clients. Coupled with the ability to collect detailed data at the granularity of individual mouse clicks, this provides a tremendous opportunity for personalizing the Web experience for clients. In e-commerce parlance this is being termed mass customization. Even outside of e-commerce, the idea of Web personalization has many applications. Recently there has been an increasing amount of research activity on various aspects of the personalization problem. Most current approaches to personalization by various Web-based companies rely heavily on human participation to collect profile information about users. This suffers from the problems of the profile data being subjective, as well getting out of date as user preferences change over time.

We have provided several techniques in which user preferences are automatically learned from Web usage data by using data mining techniques. This has the potential of eliminating subjectivity from profile data as well as keeping it updated. We have described a general architecture for automatic Web personalization based on the proposed techniques, and discussed solutions to the problems of usage data preprocessing, usage knowledge extraction, and making recommendations based on the extracted knowledge. C

References

1. Buchner, A. and Mulvenna, M.D. Discovering Internet marketing intelligence through online analytical Web usage mining. *SIGMOD Record 4*, 27 (1999).
2. 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).
3. Han, E., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., and Mobasher, B., More, J. Document categorization and query generation on the World Wide Web using WebACE. *Journal of Artificial Intelligence Review*, January 1999.
4. Han, E., Karypis, G., Kumar, V., and Mobasher, B. Clustering based on association rule hypergraphs. In *Proceedings of SIGMOD'97 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'97)*, May 1997.
5. Joachims, T., Freitag, D., Mitchell, T. WebWatcher: A tour guide for the World Wide Web. In *Proceedings of the International Joint Conference in AI (IJCAI97)*, August 1997.
6. Nasraoui, O., Frigui, H., Joshi, A., and Krishnapuram, R. Mining Web access logs using relational competitive fuzzy clustering. In *Proceedings of the Eight International Fuzzy Systems Association World Congress*, August 1999.
7. O'Conner, M. and Herlocker, J. Clustering items for collaborative filtering. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems*, Berkeley, CA, 1999.
8. Perkowitz, M. and Etzioni, O. Adaptive Web sites: automatically synthesizing Web pages. In *Proceedings of Fifteenth National Conference on Artificial Intelligence*, Madison, WI, 1998.
9. Spiliopoulou, M. and Faulstich, L.C. WUM: A Web Utilization Miner. In *Proceedings of EDBT Workshop WebDB98*, Valencia, Spain, LNCS 1590, Springer Verlag, 1999.
10. Schechter, S., Krishnan, M., and Smith, M.D. Using path profiles to predict HTTP requests. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, 1998.
11. Shardanand, U. and Maes, P. Social information filtering: algorithms for automating "word of mouth." In *Proceedings of the ACM CHI Conference*, 1995.
12. Shahabi, C., Zarkesh, A. M., Adibi, J., and Shah, V. Knowledge discovery from users Web-page navigation. In *Proceedings of Workshop on Research Issues in Data Engineering*, Birmingham, England, 1997.

Bamshad Mobasher (mobasher@cs.depaul.edu) is an assistant professor of Computer Science and the director of the Center for Web and E-Commerce Intelligence at DePaul University in Chicago, IL.

Robert Cooley (cooley@cs.umn.edu) is an independent consultant in e-commerce data analysis.

Jaideep Srivastava (srivasta@cs.umn.edu) is an associate professor of Computer Science in the Department of Computer Science and Engineering at the University of Minnesota in Minneapolis.