Integrating Web Usage and Content Mining for More Effective Personalization

Bamshad Mobasher[1], Hoghua Dai, Tao Luo, Yuqing Sun, and Jiang Zhu

School of Computer Science, Telecommunications, and Information Systems,
Depaul University, Chicago, Illinois, USA
mobasher@cs.depaul.edu

Abstract. Recent proposals have suggested Web usage mining as an enabling mechanism to overcome the problems associated with more traditional Web personalization techniques such as collaborative or content-based filtering. These problems include lack of scalability, reliance on subjective user ratings or static profiles, and the inability to capture a richer set of semantic relationships among objects (in content-based systems). Yet, usage-based personalization can be problematic when little usage data is available pertaining to some objects or when the site content changes regularly. For more effective personalization, both usage and content attributes of a site must be integrated into a Web mining framework and used by the recommendation engine in a uniform manner. In this paper we present such a framework, distinguishing between the offline tasks of data preparation and mining, and the online process of customizing Web pages based on a user's active session. We describe effective techniques based on clustering to obtain a uniform representation for both site usage and site content profiles, and we show how these profiles can be used to perform real-time personalization.

1 Introduction

The intense competition among Internet-based businesses to acquire new customers and retain the existing ones has made Web personalization an indispensable part of e-commerce. Web personalization can be defined as any action that tailors the Web experience to a particular user, or set of users. The current challenge in electronic commerce is to develop ways of gaining deep understanding into the behavior of customers based on data which is, at least in part, anonymous. We believe that the solution lies in the creation of a flexible framework and the development of new techniques for unsupervised and undirected knowledge discovery from Web usage data, and the integration of content information and meta-data with the discovered usage patterns.

Personalization based on Web usage mining has several advantages over more traditional techniques. The type of 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 relationships among Web objects based on their usage. Thus, Web usage mining will reduce the need for obtaining subjective user ratings or registration-based personal preferences. Web usage mining can also be used to enhance the effectiveness of collaborative filtering approaches [6, 16]. Collaborative filtering is often based on matching, in real-time, 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 [10], it becomes hard to scale collaborative filtering techniques to a large number of items, while maintaining reasonable prediction performance and accuracy. 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 transactions identified in the preprocessing stage.

Recent work in Web usage mining has focused on the extraction of usage patterns from Web logs for the purpose of deriving marketing intelligence [1–4, 12, 19, 20], as well as the discovery of aggregate profiles for the customization or optimization of Web sites [9, 11, 14, 17, 18]. For an up-to-date survey of Web usage mining systems see [13]. Despite the advantages, usage-based personalization can be problematic when little usage data is available pertaining to some objects or when the site content may change regularly. For more effective personalization, both usage and content attributes of a site must be integrated into a Web mining framework and used by the recommendation engine in a uniform manner.

In [7, 8], we presented a general framework for usage-based Web personalization, and proposed specific techniques based on clustering and association rule discovery to obtain dynamic recommendations from aggregate usage data. In this paper, we extend this framework to incorporate content profiles into the recommendation process as way to enhance the effectiveness of personalization actions. We discuss specific preprocessing tasks necessary for performing both content and usage mining, and present techniques based on clustering to derive aggregate profiles. Our goal is to create a uniform representation for both content and usage profiles that can be effectively used for personalization tasks by the recommendation engine in a consistent and integrated fashion. We show how the discovered knowledge can be combined with the current status of an ongoing Web activity to perform real-time personalization. Finally, we provide an experimental evaluation of the proposed techniques using real Web usage data.

2 A Web Mining Framework for Personalization

2.1 System Architecture

Figure 1 depicts a general architecture for Web personalization based on usage and content mining. The overall process is divided into two components: the offline component which is comprised of the data preparation and specific Web mining tasks, and the online component which is a real-time recommendation

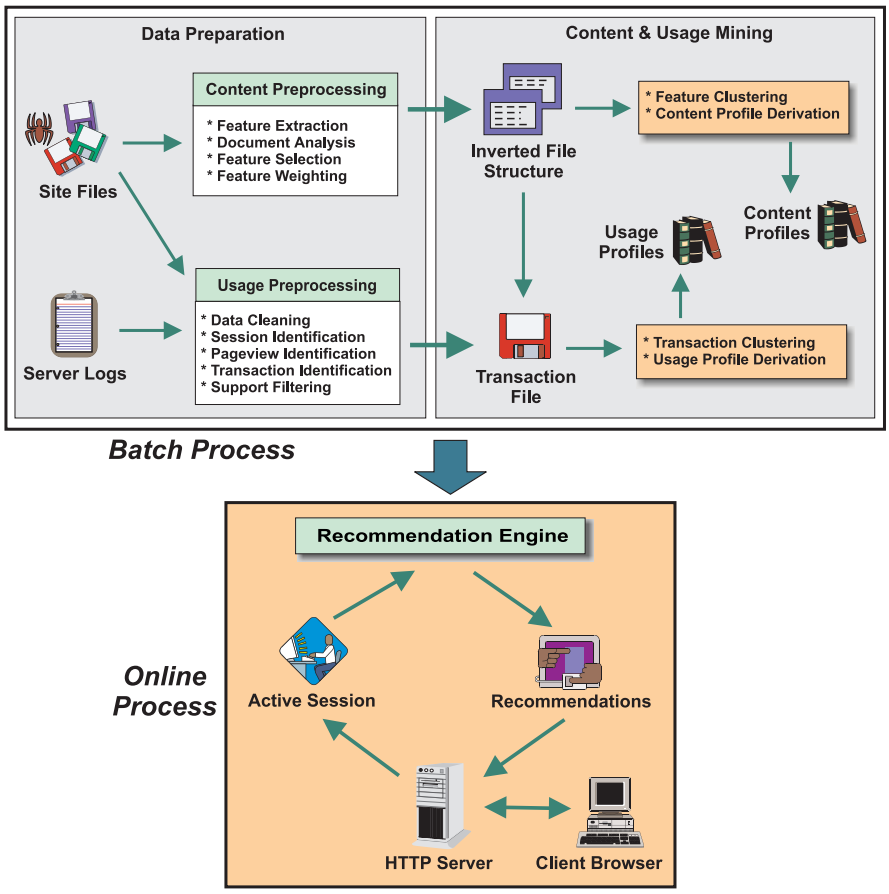

Fig. 1. A general framework for automatic personalization based on Web Mining

engine. The data preparation tasks result in aggregate structures containing the preprocessed usage and content data to be used in the mining stage. Usage mining tasks can involve the discovery of association rules, sequential patterns, pageview clusters, or transaction clusters, while content mining tasks may involve feature clustering (based on occurrence patterns of features in pageviews), pageview clustering based on content or meta-data attributes, or the discovery of (content-based) association rules among features or pageviews. In this paper, we focus on the derivation of usage profiles from transaction clusters, and the derivation of content profiles from feature clusters. In the online component of the system, the recommendation engine considers the active server session in conjunction with the discovered patterns and profiles to provide personalized content. The personalized content can take the form of recommended links or products, targeted advertisements, or text and graphics tailored to the user's perceived preferences as determined by the matching usage and content profiles.

2.2 Data Preparation for Usage and Content Mining

The required high-level tasks in usage data preprocessing are data cleaning, user identification, session identification, pageview identification, and path completion. The latter may be necessary due to client-side or proxy level caching. *User identification* is necessary for Web sites that do not use cookies or embedded session Ids. We use the heuristics proposed in [3] to identify unique user sessions form anonymous usage data and to infer cached references.

Pageview identification is the task of determining which page file accesses contribute to a single browser display. Not all pageviews are relevant for specific mining tasks. Furthermore, 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 prior domain knowledge specified by the site designer. For example, in an in an e-commerce site pageviews corresponding to product-oriented events (e.g., shopping cart changes or product information views) may be considered more significant than others. 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 URL uniquely representing the pageview), duration (for a given user session), static pageview type (e.g., content, navigational, product view, index page, etc.), and other meta-data.

Transaction identification can be performed as a final preprocessing step prior to pattern discovery in order to focus on the relevant subsets of pageviews in each user session [3]. The transaction file can be further filtered by removing very low support or very high support pageview references (i.e., references to those pageviews which do not appear in a sufficient number of transactions, or those that are present in nearly all transactions). 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 pageview references with minimal knowledge value for the purpose of personalization.

Usage preprocessing ultimately results in a set of n pageview records appearing in the transaction file, $P = \{p_1, p_2, \cdots, p_n\}$, with each pageview record uniquely represented by its associated URL, and a set of m user transactions, $T = \{t_1, t_2, \cdots, t_m\}$, where each $t_i \in T$ is a subset of P. To facilitate various data mining operations such as clustering, we view each transaction t as an n-dimensional vector over the space of pageview references, $t = \langle w(p_1, t), w(p_2, t), \cdots, w(p_n, t)\rangle$, where $w(p_i, t)$ is a weight, in the transaction t, associated with the pageview represented by $p_i \in P$. The weights can be determined in a number of ways, for example, binary weights can be used to represent existence or non-existence of a product-purchase or a documents access in the transaction. On the other hand, the weights can be a function of the duration of the associated pageview in order to capture the user's interest in a content page. The weights may also, in part, be based on domain-specific significance weights assigned by the analyst.

Content preprocessing involves the extraction of relevant features from text and meta-data. Meta-data extraction becomes particularly important when dealing with product-oriented pageviews or those involving non-textual content. In

the current implementation of our framework features are extracted from meta-data embedded into files in the form of XML or HTML meta-tags, as well as from the textual content of pages. In order to use features in similarity computations, appropriate weights must be associated with them. For features extracted from meta-data, we assume that feature weights are provided as part of the domain knowledge specified by the site designer. For features extracted from text we use a standard function of the term frequency and inverse document frequency (tf.idf) for feature weights as commonly used in information retrieval [5, 15].

Specifically, each pageview p is represented as a k-dimensional feature vector, where k is the total number of extracted features from the site in a global dictionary. Each dimension in a feature vector represents the corresponding feature weight within the pageview. Thus, the feature vector for a pageview p is given by: $p = \langle fw(p, f_1), fw(p, f_2), \cdots, fw(p, f_k) \rangle$ where $fw(p, f_j)$, is the weight of the jth feature in pageview $p \in P$, for $1 \leq j \leq k$. For features extracted from textual content of pages, the feature weight is obtained as the normalized tf.idf value for the term. Finally, in order to combine feature weights from meta-data (specified externally) and feature weights from the text content, proper normalization of those weights must be performed as part of preprocessing. The feature vectors obtained in this way are organized into an inverted file structure containing a dictionary of all extracted features and posting files for each feature specifying the pageviews in which the feature occurs along with its weight. Conceptually, this structure can be viewed as a feature-pageview matrix in which each column is a feature vector corresponding to a pageview.

2.3 Discovery of Aggregate Usage Profiles

The transaction file obtained in the data preparation stage can be used as the input to a variety of data mining algorithms. 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 (i.e., actionable) "aggregate profiles" from these patterns. Ideally, a profile captures an aggregate view of the behavior of subsets of users based their common interests or information needs. In particular, aggregate profiles must be able to 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. Furthermore, they should provide the capability to distinguish among pageviews in terms of their significance within the profile.

Based on these requirements, we have found that representing usage profiles as weighted collections of pageview records provides a great deal of flexibility. Each item in a usage profile is a URL representing a relevant pageview, and can have an associated weight representing its significance within the profile. The profiles can be viewed as ordered collections (if the goal is to capture the navigational path profiles followed by users [12]), or as unordered (if the focus is on capturing associations among specified content or product pages). This uniform representation allows for the recommendation engine to easily integrate different kinds of profiles (i.e., content and usage profiles, as well as multiple profiles

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

Given the mapping of user transactions into a multi-dimensional space as vectors of pageview, standard clustering algorithms, such as k-means, generally partition this space into groups of transactions that are close to each other based on a measure of distance or similarity. Such a clustering will result in a set $TC = \{c_1, c_2, \cdots, c_k\}$ of clusters, where each c_i is a subset of the set of transactions T. Ideally, each cluster represents a group of users with similar navigational patterns. However, transaction clusters by themselves are not an effective means of capturing an aggregated view of common user profiles. Each transaction cluster may potentially contain thousands of user transactions involving hundreds of pageview references. Our ultimate goal in clustering user transactions is to reduce these clusters into weighted collections of pageviews which represent aggregate profiles.

An effective method for the derivation of profiles from transaction clusters was first proposed in [8]. For each transaction cluster $c \in TC$, we compute the mean vector m_c. The mean value for each pageview in the mean vector is computed by finding the ratio of the sum of the pageview weights across transactions in c to the total number of transactions in the cluster. The weight of each pageview within a profile is a function of this quantity thus obtained. In generating the usage profiles, the weights are normalized so that the maximum weight in each usage profile is 1, and low-support pageviews (i.e. those with mean value below a certain threshold μ) are filtered out. Thus, given a transaction cluster c, we construct a usage profile pr_c as a set of pageview-weight pairs:

$$pr_c = \{\langle p, weight(p, pr_c)\rangle \mid p \in P, weight(p, pr_c) \geq \mu\}$$

where the significance weight, $weight(p, pr_c)$, of the pageview p within the usage profile pr_c is given by:

$$weight(p, pr_c) = \frac{1}{|c|} \cdot \sum_{t \in c} w(p, t)$$

and $w(p, t)$ is the weight of pageview p in transaction $t \in c$. Each profile, in turn, can be represented as vectors in the original n-dimensional space.

2.4 Discovery of Content Profiles

We use precisely the same representation for content profiles (i.e., as a weighted collection of pageviews). In contrast to usage profiles, content profiles represent different ways pages with partly similar content may be grouped together. Our goal here is to capture common interests of users in a group of pages because specific portions of their contents are similar. Different groups of users may be interested in different segments of each page, thus content profiles must capture overlapping interests of users.

Clusters of pageviews obtained using standard clustering algorithms which partition the data are not appropriate as candidates for content profiles. To obtain content profiles, instead of clustering pageviews (as k-dimensional feature vectors, where k is the number of extracted features in the global site dictionary), we cluster the features. Using the inverted feature-pageview matrix obtained in the content preprocessing stage, each feature can be viewed as an n-dimensional vector over the original space of pageviews. Thus, each dimension in the pageview vector for a feature is the weight associated with that feature in the corresponding pageview. We use multivariate k-means clustering technique to cluster these pageview vectors. Now, given a feature cluster G, we construct a content profile C_G as a set of pageview-weight pairs:

$$C_G = \{\langle p, weight(p, C_G)\rangle \mid p \in P,\ weight(p, C_G) \geq \tau\}$$

where the significance weight, $weight(p, C_G)$, of the pageview p within the content profile is obtained as follows:

$$weight(p, C_G) = \frac{\sum_{f \in G} fw(p, f)}{\sum_{i=1}^{n} \sum_{f \in G} fw(p_i, f)}$$

and $fw(p, f)$ is the weight of a feature f in pageview p. As in the case of usage profiles, we normalize pageview weights so that the maximum weight in each profile is 1, and we filter out pageviews whose weight is below a specified significance threshold, τ. Note that the representation of content profiles as a set of pageview-weight pairs is identical to that for usage profiles discussed earlier. This uniform representation allows us to easily integrate both types of profiles with the recommendation engine.

3 Integrating Content and Usage Profiles for Personalization

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 the objects (links, ads, text, products, etc.) that most closely match the current user profile. The essential aspect of computing a recommendation set for a user is the matching of current user's activity against aggregate usage profiles. The recommended objects are added to the last page in the active session accessed by the user before that page is sent to the browser. 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 sub-sessions have a length of no more than 2 or 3 references. We capture the user history depth within a sliding window over the current session. The sliding window of size n over the active session allows only the last n visited pages to influence the recommendation value of items

in the recommendation set. Finally, the structural characteristics of the site or prior domain knowledge can also be used to associate an additional measure of significance with each pageview in the user's active session.

In our proposed architecture, both content and usage profiles are represented as sets of pageview-weight pairs. This will allow for both the active session and the profiles to be treated as n-dimensional vectors over the space of pageviews in the site. Thus, given a content or usage profile C, we can represent C as a vector $C = \langle w_1^C, w_2^C, \cdots, w_n^C \rangle$, where

$$w_i^C = \begin{cases} weight(p_i, C), & \text{if } p_i \in C \\ 0, & \text{otherwise} \end{cases}$$

Similarly, the current active session S is also represented as a vector $S = \langle s_1, s_2, ..., s_n \rangle$, where s_i is a significance weight associated with the corresponding pageview reference, if the user has accessed p_i in this session, and $s_i = 0$, otherwise. We can compute the profile matching score using a similarity function such as the normalized cosine measure for vectors:

$$match(S, C) = \frac{\sum_k w_k^C \cdot S_k}{\sqrt{\sum_k (S_k)^2 \times \sum_k (w_k^C)^2}}.$$

Note that the matching score is normalized 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 a larger cluster representing a user profile. Given a profile C and an active session S, a recommendation score, $Rec(S, p)$, is computed for each pageview p in C as follows:

$$Rec(S, p) = \sqrt{weight(p, C) \ \cdot \ match(S, C)}.$$

If the pageview p is in the current active session, then its recommendation value is set to zero. We obtain the usage recommendation set, $UREC(S)$, for current active session S by collecting from each usage profile all pageviews whose recommendation score satisfies a minimum recommendation threshold ρ, i.e.,

$$UREC(S) = \{w_i^C \mid C \in UP, \text{and } Rec(s, w_i^C) \geq \rho\},$$

where UP is the collection of all usage profiles. Furthermore, for each pageview that is contributed by several usage profiles, we use its maximal recommendation score from all of the contributing profiles. In a similar manner, we can obtain the content recommendation set $CREC(S)$ from content profiles. Different methods can be used for combining the two recommendation sets depending on the goals of personalization and the requirements of the site. In our case, for each pageview we take the maximum recommendation value across the two recommendation sets. This allows, for example, content profiles to contribute to the recommendation set even if no matching usage profile is available and vice versa.

4 Experimental Results

We conducted a series of experiments with real usage data from the site for the newsletter of the *Association for Consumer Research* (from July 1998 to June 1999). The site contains a variety of news items, including President's columns, conference announcements, and call-for-papers for a number of conferences and journals. The usage preprocessing steps described earlier resulted in a user transaction file containing 18430 user transactions with a total of 62 pageviews represented uniquely by their associated URLs. The transaction clustering process yielded 16 transaction clusters representing different types of user access patterns. A threshold of 0.5 was used to derive usage profiles from transaction clusters (i.e., profiles contained only those pageviews appearing in at least 50% of transactions). In the content preprocessing stage, a total of 566 significant features were extracted with each document contributing at most 20 significant features to the global dictionary (normalized term frequency was used for measuring feature significance). Feature clustering using multivariate k-means resulted in 28 feature clusters from which the corresponding (overlapping) content profiles were derived.

Figure 2 depicts an example of two overlapping content profiles. The top significant features where listed for each document for illustrative purposes. These features indicated why each of the pageviews were included in each content profile. The first profile captures those documents in which a portion of the content relates to global and international business management and marketing. On the other hand, the second profile includes documents about consumer behavior and psychology in marketing. Note that documents which contain content related to both topics have been included in both profiles. Usage profiles are represented in the same manner, but they capture overlapping aggregate usage patterns of the site users.

Weight	Pageview ID	Significant Features (stems)
1.00	CFP: One World One Market	world challeng busi co manag global
0.63	CFP: Int'l Conf. on Marketing & Development	challeng co contact develop intern
0.35	CFP: Journal of Global Marketing	busi global
0.32	CFP: Journal of Consumer Psychology	busi manag global
Weight	**Pageview ID**	**Significant Features (stems)**
1.00	CFP: Journal of Psych. & Marketing	psychologi consum special market
1.00	CFP: Journal of Consumer Psychology I	psychologi journal consum special market
0.72	CFP: Journal of Global Marketing	journal special market
0.61	CFP: Journal of Consumer Psychology II	psychologi journal consum special
0.50	CFP: Society for Consumer Psychology	psychologi consum special
0.50	CFP: Conf. on Gender, Market., Consumer Behavior	journal consum market

Fig. 2. Two Overlapping Content Profiles

The recommendation engine was used for a sample user session using a window size of 2. Figure 3 shows the system recommendations based only on usage profiles, while Figure 2 shows the results from only the content profiles. In these

tables, the first column shows the pageviews contained in the current active session. The last pageview in each session window represents the current location of the user in the site. The right-hand column gives the recommendation score obtained using the techniques discussed in the previous section. It is clear from these examples that the combination of recommendations from both content and usage profiles can provide added value to the user. For example, in the usage-based recommendations, the user's visit to "`ACR Board of Directors Meeting`" did not yield any recommendations with a score above the specified threshold (0.5), while content-based recommendations produced some pages with related content. On the other hand, navigating to the page "`Conference Update`" resulted in content-based recommendations for pages with only general information and news about conferences, while usage profiles yielded a number of specific recommendations that the site users interested in conferences and calls for papers tend visit.

Pages in Active Session Window	Recommendations	Score
* ACR Board of Directors Meeting	NO RECOMMENDATIONS	
* ACR Board of Directors Meeting * Conference Update	ACR 1999 Annual Conference	0.57
	CFP: ACR'99 Asia-Pacific Conf.	0.53
	CFP: Int'l Conf. on Marketing & Development	0.52
	CFP: ACR'99 European Conf.	0.51
	President's Column - December, 1998	0.50
* Conference Update * CFP: Journal of Psych. & Marketing	ACR News Special Topics	0.64
	ACR 1999 Annual Conference	0.57
	CFP: Int'l Conf. on Marketing & Development	0.53
	CFP: ACR'99 European Conf.	0.53
	CFP: Winter 2000 SCP Conference	0.52
	CFP: Int'l Research Seminar in Marketing	0.50
* CFP: Journal of Psych. & Marketing * CFP: Journal of Global Marketing	ACR News Special Topics	0.59

Fig. 3. Recommendations Based on Usage Profiles

5 Conclusions and Future Work

We have presented a general framework for personalization based on usage and content mining, in which the user preference is automatically learned from Web usage data and integrated with domain knowledge and the site content. This has the potential of eliminating subjectivity from profile data as well as keeping it up-to-date. Furthermore, the integration of usage and content mining increases the usefulness and accuracy of the resulting recommendations. Our experimental results indicate that the techniques discussed here are promising, each with its own unique characteristics, and bear further investigation and development. Our future work in this area will include automatic classification of pageview

Pages in Active Session Window	Recommendations	Score
* ACR Board of Directors Meeting	Special Topics - ACR Letters and Special Topics	0.70
	Special Topics - ACR Board of Directors Agenda	0.55
	Special Topics - ACR Appointments	0.52
* ACR Board of Directors Meeting	Call for Papers	0.67
* Conference Update	ACR News Updates	0.54
	ACR News Special Topics	0.51
* Conference Update	Call for Papers	0.67
* CFP: Journal of Psych. & Marketing	CFP: Journal of Consumer Psych. I	0.59
	CFP: Journal of Psych. & Marketing	0.56
	CFP: Journal of Consumer Psych. II	0.51
	CFP: Journal of Global Marketing	0.50
* CFP: Journal of Psych. & Marketing	CFP: Journal of Consumer Psych. I	0.77
* CFP: Journal of Global Marketing	CFP: Journal of Psych. & Marketing	0.73
	CFP: Journal of Consumer Psych. II	0.59
	CFP: Marketing & Public Policy Conf.	0.56
	CFP: Conf. on Gender, Marketing , Consumer Behavior	0.54

Fig. 4. Recommendations Based on Content Profiles

types and better integration of a variety of pageview types (such as those representing different e-commerce or product-oriented events) into the mining and recommendation process.

References

1. A. Buchner and M. D. Mulvenna. Discovering internet marketing intelligence through online analytical Web usage mining. *SIGMOD Record*, (4) 27, 1999.
2. R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *International Conference on Tools with Artificial Intelligence*, pages 558–567, Newport Beach, 1997. IEEE.
3. R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining World Wide Web browsing patterns. *Journal of Knowledge and Information Systems*, (1) 1, 1999.
4. M. S. Chen, J. S. Park, and P. S. Yu. Data mining for path traversal patterns in a Web environment. In *Proceedings of 16th International Conference on Distributed Computing Systems*, 1996.
5. W. B. Frakes, R. Baeza-Yates. *Information Retrieval Data Structures and Algorithms*. Prentice Hall, Englewood Cliffs, NJ, 1992.
6. J. Herlocker, J. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*, August 1999.
7. B. Mobasher, R. Cooley, and J. Srivastava. Creating adaptive web sites through usage-based clustering of urls. In *IEEE Knowledge and Data Engineering Workshop (KDEX'99)*, November 1999.
8. B. Mobasher. A Web personalization engine based on user transaction clustering. In *Proceedings of the 9th Workshop on Information Technologies and Systems (WITS'99)*, December 1999.
9. O. Nasraoui, H. Frigui, A. Joshi, R. Krishnapuram. Mining Web access logs using relational competitive fuzzy clustering. In *Proceedings of the Eight International Fuzzy Systems Association World Congress*, August 1999.

10. M. O'Conner, J. Herlocker. Clustering items for collaborative filtering. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems*, Berkeley, CA, 1999.
11. M. Perkowitz and O. Etzioni. Adaptive Web sites: automatically synthesizing Web pages. In *Proceedings of Fifteenth National Conference on Artificial Intelligence*, Madison, WI, 1998.
12. M. Spiliopoulou and L. C. Faulstich. WUM: A Web Utilization Miner. In *Proceedings of EDBT Workshop WebDB98*, Valencia, Spain, LNCS 1590, Springer Verlag, 1999.
13. J. Srivastava, R. Cooley, M. Deshpande, P-T. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations*, (1) 2, 2000.
14. S. Schechter, M. Krishnan, and M. D. Smith. Using path profiles to predict HTTP requests. In *Proceedings of 7th International World Wide Web Conference*, Brisbane, Australia, 1998.
15. G. Salton, M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
16. U. Shardanand, P. Maes. Social information filtering: algorithms for automating "word of mouth." In *Proceedings of the ACM CHI Conference*, 1995.
17. C. Shahabi, A. Zarkesh, J. Adibi, and V. Shah. Knowledge discovery from users Web-page navigation. In *Proceedings of Workshop on Research Issues in Data Engineering*, Birmingham, England, 1997.
18. T. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal. From user access patterns to dynamic hypertext linking. In *Proceedings of the 5th International World Wide Web Conference*, Paris, France, 1996.
19. K. Wu, P. S. Yu, and A. Ballman. Speedtracer: A web usage mining and analysis tool. *IBM Systems Journal*, 37(1), 1998.
20. O. R. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying olap and data mining technology on web logs. In *Advances in Digital Libraries*, pages 19-29, Santa Barbara, CA, 1998.