Web User Segmentation Based on a Mixture of Factor Analyzers

Yanzan Kevin Zhou[1] and Bamshad Mobasher[2]

[1] eBay Inc., San Jose, CA
yanzzhou@ebay.com
[2] DePaul University, Chicago, IL
mobasher@cs.depaul.edu

Abstract. This paper proposes an approach for Web user segmentation and online behavior analysis based on a mixture of factor analyzers (MFA). In our proposed framework, we model users' shared interests as a set of common latent factors extracted through factor analysis, and we discover user segments based on the posterior component distribution of a finite mixture model. This allows us to measure the relationships between users' unobserved conceptual interests and their observed navigational behavior in a principled probabilistic manner. Our experimental results show that the MFA-based approach results in finer-grained representation of user behavior and can successfully discover heterogeneous user segments and characterize these segments with respect to their common preferences.

1 Introduction

Web sites are increasingly becoming more complex, often involving a myriad of functions and tasks that can be performed online, or diverse content areas that span a variety of topics or subtopics. Therefore, increasingly sophisticated models are necessary to precisely capture Web user's interests and preferences. *Web usage mining* [1,10,6] plays a key role in Web user modeling. It is the most direct approach to studying Web users' online behavior since its primary data source is clickstream data, which is generated by Web users' interaction with a Web site and recorded in application or Web server log files. A variety of data mining and statistical techniques have been applied to discover useful patterns, such as Web page association rules [7,4] or user clusters [8].

In particular, user segmentation is a widely used approach for characterizing and understanding user behavior and interests in a Web site. Both distance-based clustering methods, such as k-means, and probabilistic density-based mixture models can be used for market segmentation purposes [11]. Clustering methods find groups of data objects based on their distances or distribution similarity. A disadvantage of distance-based methods is that no probabilistic inference can be made and, for high-dimensional data, the distance computation can be prone to noise and outliers. This can be partially remedied by model-based clustering,

K. Bauknecht et al. (Eds.): EC-Web 2006, LNCS 4082, pp. 11–20, 2006.

in which a mixture of probabilistic distributions are assumed to have generated the data, and the data objects that follow the same distribution can be regarded as a cluster. However, ordinary mixture models such as a Mixture of Gaussians (MoG) still suffer parameter over-fitting problems emanating from high-dimensional feature space. Furthermore, standard mixture models cannot discover the latent dimensional structure of the observed data which can "explain" the relationships among data objects.

Well-established latent variable models, such as Principal Component Analysis (PCA) and Factor Analysis (FA), are generally used for dimensionality reduction and discovery of latent structures in data. While the commonly used PCA method assumes zero-noise model, FA-based approaches distinguish between common variance and noise variance for each of the observed variables. Such noise discrimination effect is particularly important in the typically noisy Web navigation data [12].

In this paper we propose an approach for Web user segmentation and online behavior analysis based on a *Mixture of Factor Analyzers* (MFA). MFA is natural integration of finite mixture models and factor analysis, resulting in a statistical method which concurrently performs clustering and, within each cluster, local dimensionality reduction. This presents several benefits over approaches in which clustering and dimensionality reduction are performed separately. First, different features may be correlated within different clusters and thus the metric for dimensionality reduction may need to vary between different clusters. Conversely, the metric induced in dimensionality reduction may guide the process of cluster formation, i.e. different clusters may appear more separated depending on the local metric [3].

MFA has been shown effective in simultaneous case-space clustering and feature-space dimensionality reduction for the purpose of speech recognition and face detection [9,2]. However, to the best of our knowledge, MFA has not been used in modeling of Web user navigational patterns. Web usage data tends to be high dimensional, and Web users generally have different navigational behaviors based on their intended tasks or information needs, however, with common subpatterns, resulting in noisy patterns corresponding to multiple modalities. This, we believe, makes MFA particularly useful in Web user modeling: the mixture component variables model the global variation among individual Web users, and the latent dimensions in the factor analysis model allow for the conceptual representation of user's hidden interests without the typical noise. The discovered patterns are not only useful for online user behavior understanding, but also for other important e-commerce applications such as collaborative recommendation.

The paper is organized as follows. In Section 2 we discuss our MFA-based approach to model the multimodal Web navigation data and quantify the relationship between Web users' latent interests, segment memberships, and their observed behavior in user sessions. In Section 3 we introduce our approach for user segmentation based on posterior latent variable distribution in MFA. Section 4 presents our empirical results and verification based on experimental study of online user behavior on real world Web usage data.

2 Mixture of Factor Analyzers for Web Usage Data

We assume that appropriate preprocessing (such as data cleaning, sessionization, spider removal, etc.) has been performed on raw Web server logs [1], resulting in a set of p pages $\mathcal{D} = \{D_1, D_2, \cdots, D_p\}$ and a set of n user sessions $\mathcal{U} = \{U_1, U_2, \cdots, U_n\}$. Each user session can be represented as a p-dimensional vector $\mathbf{u} \in \mathbb{R}^p$ as $\mathbf{u} = [d_1, d_2, \cdots, d_p]^T$, representing user-page observations in that session, where the value of d_j is a weight associated with page D_j in the session (in our experiments, the weights are a function of the time spent on each page during the session).

In this section, we first present the basic elements of factor analysis for modeling Web navigational patterns, and then, we present our framework for extending the standard factor analysis model to a mixture model.

2.1 Using Factor Analysis to Model Web Usage Patterns

In standard maximum likelihood factor analysis (FA), an n-dimensional real-valued data vector $\mathbf{u}$ (in our case, a user session) is modeled using a k-dimensional vector of real-valued factors, $\mathbf{z}$, where k is generally much smaller than n. Since, in usage data these factors closely correspond to aggregate common interests of users, we call $\mathbf{z}$ a *preference vector*. Specifically, given $\mathbf{z} = [z_1, z_2, \cdots, z_k]^T \in \mathbb{R}^k$, we can view a user's access to a page D_i during a session as the sum of combined "influences" by the set of latent variables, each representing an abstract common preference. In other words, $d_i = l_{i1}z_1 + l_{i2}z_2 + ... + l_{ik}z_k + \epsilon_i$, where coefficient l_{ij} indicates how strongly the page D_i is related to user's preference Z_j, and ϵ_i represents independent random variance (noise) that is not accounted for by the k latent variables (which account for common variances). Since $k \ll p$, the original high dimensional data (at the page level) is mapped to a much lower-dimensional latent space with unwanted information modeled as random noise.

We can arrive at the density-based factor analysis model by defining proper probabilistic density functions (PDF) over the latent variables, and assuming a generative model for Web user session observations:

$$p(\mathbf{u}) = \int_{\mathbf{z}} p(\mathbf{u}|\mathbf{z})p(\mathbf{z})d\mathbf{z}, \quad \mathbf{z} \in \mathbb{R}^k. \tag{1}$$

We also obtain a linear Gaussian factor analysis model with the assumptions of a multivariate Gaussian prior over the latent variables $\mathbf{z}$, and an independent Gaussian noise model over ϵ, i.e.,

$$\mathbf{u} = \mathbf{L}\mathbf{z} + \epsilon; \quad \mathbf{z} \sim \mathcal{N}_k(0, \mathbf{I}), \quad \epsilon \sim \mathcal{N}_p(0, \mathbf{\Psi}), \tag{2}$$

where $\sim \mathcal{N}_k(0, \mathbf{I})$ denotes a k-variate joint Gaussian with zero mean and identity covariance matrix, and $\mathbf{\Psi}$ is the diagonal covariance matrix of random variances $diag(\sigma_i^2)$. Then, based on Equation 1 we can derive the unconditional PDF for the data, $p(\mathbf{u}) = \mathcal{N}_p(0, \mathbf{L}\mathbf{L}^T + \mathbf{\Psi})$.

2.2 MFA-Based Web User Modeling

Conceptually, We regard a user's online navigation behavior as the following generative process:

1. When a user u comes to a Web site, the user is assigned to a certain activity group g based on his current browsing history.
2. The expected activity of the user in group g is then determined by his k-dimensional preference vector.

Mathematically, a mixture model is a linear combination of m mixture component densities weighted by their prior probabilities (mixing proportion), i.e.,

$$p(\mathbf{u}) = \sum_{g=1}^{m} p(\mathbf{u}, g) = \sum_{g=1}^{m} p(\mathbf{u}|g)P(g) \tag{3}$$

where g is a discrete mixture variable taking value $g \in \{1, 2, \cdots, m\}$ and satisfying the condition $\sum_{g=1}^{m} P(g) = 1, P(g) \geq 0$.

If in the mixture model (Equation 3), each class-conditional density is a latent variable model (equation 1), then we obtain a *mixture of latent variable models*. In our case, we obtain a mixture of factor analyzers. The marginal observation distribution is obtained by integrating over both discrete mixture variable g and continuous latent variables $\mathbf{z}$

$$p(\mathbf{u}) = \sum_{g=1}^{m} \int_{\mathbf{z}} p(\mathbf{u}|\mathbf{z}, g)p(\mathbf{z})P(g)d\mathbf{z} \tag{4}$$

resulting in the following concrete unconditional PDF for MFA:

$$p(\mathbf{u}) = \sum_{g=1}^{m} \mathcal{N}_p(\mu_g, \mathbf{L}_g\mathbf{L}_g^T + \mathbf{\Psi})P(g) \tag{5}$$

where each mixture component has its own mean μ_g and loading matrix $\mathbf{L}_g$ together with its prior probability $P(g)$. This allows each factor analyzer to model the data covariance structure in a different part of the input space. The estimation of MFA parameters $\mu_g, \mathbf{L}_g, \mathbf{\Psi}_g, \{P(g)|g \in [1, m]\}$, is achieved by the EM algorithms, which is commonly used for latent variable models. Interested readers can refer to [3] and [5] for more details.

MFA is a nonlinear extension of linear Gaussian FA addressing both global data modal heterogeneity and local dimensionality reduction, thus combining both FA and Mixture model's merits which is particularly desirable for Web data.

3 MFA-Based User Segmentation

Since each mixture component models a subpopulation of Web users following the same distribution, we can naturally derive user segments based on the relationship between users and components:

$$P(g|\mathbf{u}) \propto P(g)p(\mathbf{u}|g) = P(g)\mathcal{N}_p(\mu_g, \mathbf{L}_g\mathbf{L}_g^T + \mathbf{\Psi}). \quad (6)$$

In our approach, for each of the m mixture components, we choose those users whose posterior memberships are greater than a certain threshold as its representative users. This is, essentially, a soft clustering of user sessions based on membership probabilities given in Equation 6. Hard clustering can also be achieved by simply allocating each user into only one of the segments which satisfies $argmax_g P(g|\mathbf{u})$. Such a set of x users in a segment, $U^g = \{\mathbf{u}_1^g, \mathbf{u_2}^g, \cdots, \mathbf{u}_x^g\}$, would have similar preference patterns (determined by the shared loading matrix $\mathbf{L}_g$).

In addition to user segment derivation, expected preference values for each user segment can be further derived according to Equation (7). That is, the conditional expectations of a user's preference values is obtained as the factor scores associated with a certain mixture component:

$$\begin{aligned} E(\mathbf{z}, g|\mathbf{u}) &= \int_{-\infty}^{+\infty} \mathbf{z}p(\mathbf{z}, g|\mathbf{u})d\mathbf{z} = P(g|\mathbf{u}) \int_{-\infty}^{+\infty} \mathbf{z}p(\mathbf{z}|\mathbf{u}, g)d\mathbf{z} \\ &= P(g|\mathbf{u})E[\mathbf{z}|\mathbf{u}, g] \end{aligned} \quad (7)$$

where $E[\mathbf{z}|\mathbf{u}, g] = \mathbf{L}_j^T(\mathbf{L}_j\mathbf{L}_j^T + \mathbf{\Psi})^{-1}(\mathbf{u} - \mu_j)$, and $P(g|\mathbf{u})$ is in Equation (6). Based on these preference values we can easily identify the segment's dominant factors which characterize the behavior of users within that segment.

To create an aggregate representation of the user segment U^g, we compute the weighted centroid of all the observation vectors in the segment (weighted by the membership), which results in a representation of the segment as a set of page-weight pairs. The algorithm for generating the aggregate representation of each user segment is based on the following two steps:

1. For each mixture component g, choose those user sessions with posterior memberships $argmax_g P(g|\mathbf{u})$ to form a candidate session set U^g, where $P(g|\mathbf{u})$ is determined by the posterior probability as in Equation 6.
2. Recall that each user session $\mathbf{u}_i \in U^g$ is a p-variate page vector. For each set U^g, compute its weighted centroid vector of pages as

$$\mathbf{v}^g = \frac{1}{|U^g|} \sum_{\mathbf{u}_i \in U^g} [\mathbf{u}_i \times P(g|\mathbf{u}_i)],$$

where $|U^g|$ denotes the total number of sessions in set U^g.

Therefor for each user segment g, we derive a centroid-based user model represented as a page vector $\mathbf{v}^g$.

4 Experiments and Evaluation

In this section we evaluate the MFA-based model fit and the predictive effectiveness of our derived segments using two different data sets: CTI data and ACR

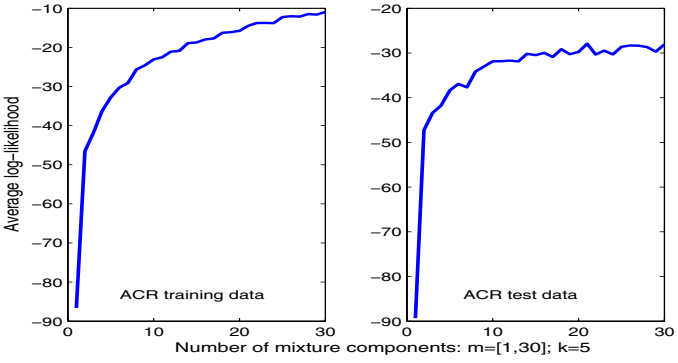

Fig. 1. MFA Log-likelihood vs. number of mixture components in ACR data

data. CTI data set is based on the server logs of the host Computer Science department spanning a one-month period containing 21,299 user sessions and 692 Web pages after preprocessing. The site is highly dynamic, involving numerous online applications, advising, faculty-specific Intranet applications, etc. Thus, we expect the discovered usage patterns to reflect various functional tasks performed by diverse groups of users. The ACR data set is based on the Web server log files of Association for Consumer Research Web site which contained 7,834 user sessions and 40 Web pages after preprocessing. Each data set was randomly divided into multiple training and test sets for the purpose of cross-validation.

4.1 Evaluation of Model Fit with Average Log-Likelihood

Average log-likelihood on training data measures the goodness of fit of the model to the training data, i.e., the likelihood that an observed case in the data is generated by the model (represented by its parameter set Θ). Specifically, the average log-likelihood $\bar{\mathcal{L}}(\Theta|data) = \frac{1}{n} \ln \prod_{i=1}^{n} p(\mathbf{u}_i|\Theta)$ of MFA is computed as follows.

$$\bar{\mathcal{L}} = \frac{1}{n} \sum_{i=1}^{n} \ln \sum_{g=1}^{m} P(g) \mathcal{N}_p(\mathbf{u}_i|\mu_g, \mathbf{L}_g \mathbf{L}_g^T + \mathbf{\Psi}); \tag{8}$$

We should expect that the log-likelihood on the evaluation data is consistent with the training data, in order to obtain a relatively generalizable model. An example on ACR data is illustrated in Figure 1. First, we fit MFA to the training data, and vary the number of mixture components from $m \in [1, 30]$. Note that $m = 1$ corresponds to the standard single Factor Analysis. Then we evaluate the learned models on the evaluation data set based on their average log-likelihoods. Each likelihood value is an average of 5 runs of MFA EM algorithm to offset the random model initialization effect. From the training data likelihood (left) we see a significant improvement of the average log-likelihood from single FA

($m = 1$) to a mixture of two factor analyzers ($m = 2$) and a slower rate of increase thereafter. This tells us that the usage data can be better modeled by MFA than FA to address multimodal distribution problems.

Furthermore, from the test set likelihood curve (right) we can roughly tell when the model starts to over-fit. In this particular case, we see the average log-likelihood curve starts to level off with the number of mixture components $m = 10$. We found that this kind of cross-validation is effective for selecting a desirable number of mixture components although other methods such as BIC may also be applied.

4.2 Analysis of User Segment Behavior

We have conducted another set of experiments specifically to verify the following hypotheses:

- **Hypothesis 1:** Each distribution component of the mixture model MFA corresponds to a finer-grained representation of a group of users sharing similar navigation behavior pattern than an ordinary FA does since mixture models can capture mixed underlying distributions behind the data;
- **Hypothesis 2:** Each particular component of MFA reflects an activity type of a group of users with common interests, which are represented by the corresponding dominant latent factors.

In the CTI data, we manually identified three significant activity types, namely "faculty advising","graduate application" and "discussion forums". For each activity type, we isolated its corresponding user sessions resulting in three separate data sets as evaluation data. As shown in Figure 2, for factor analysis, the average log-likelihoods on the three evaluation sets are all smaller than MFA (column *mixture*). We have also found that MFA has higher likelihood value in the training data. This is not surprising since based on our experience, general Web usage data present multimodal characteristics with mixed underlying distributions.

In order to verify our hypotheses, we carried out further experiments by separately evaluating each of the individual mixture component models in MFA. As we know that each mixture component g of MFA has a corresponding parameter set. If at least one of these components can capture a distinct user behavior type better than FA, we should expect a higher likelihood when evaluated on that behavior type data than FA which models the entire training data with mixed types. Thus we applied these individual component models to all three evaluation sets to obtain average log-likelihoods respectively as shown in Figure 2 (columns indicated by *comp.*). We can see that in the segment of "faculty advising", the component $g = 5$ has the highest likelihood (-347.72) among the five individual component models and also higher than FA (-363.65). This kind of observation is consistent on all three evaluation data sets including "graduate application" and "discussion forums" data. We further observe that although the best single component with the highest likelihood better captures a distinct

User Segments	FA ($k = 10$)	MFA ($m = 5$, $k = 10$)					
		$g = 1$ (comp.)	$g = 2$ (comp.)	$g = 3$ (comp.)	$g = 4$ (comp.)	$g = 5$ (comp.)	$g = [1,5]$ (mixture)
Faculty advising	-363.65	-415.25	-428.91	-395.68	-393.12	**-347.72**	-332.82
graduate application	-317.25	**-307.33**	-390.21	-341.94	-349.97	-377.64	-293.09
Discussion Forums	-360.88	-389.63	-388.86	-376.14	**-348.12**	-387.43	-309.61

Fig. 2. Likelihood comparison of FA, MFA and its individual component models

behavior type than standard FA, the combined mixture model MFA, which incorporates all the components enjoys the highest likelihoods than either FA or individual components on all the evaluation data sets. Intuitively, this is because Web user segments generally represent diversified activities while having similar dominant navigation interests, which is better captured by a combination of mixture model and factor analysis model.

It would be interesting to take a closer look at the internal structure of a mixture component model. Since each component is essentially a factor model, we want to verify whether the dominant factors in a group of users assigned to component g would match their dominant activity type. For example, from Figure 2 we know that for the evaluation data of user segment "graduate application", the best matched mixture component model is $g = 1$ ($\bar{\mathcal{L}} = -317.25$). In other words, users having been assigned to component $g = 1$ should have the dominant interest of "graduate application". In order to verify this, we selected user sessions from the training set whose membership probability equals to $\arg\max_g P[g|\mathbf{u}, g]$ and computed their mean preference scores on five latent factors. We found that the dominant factor of the segment associated with $g = 1$, which is "graduate application" as we have known, does have the largest mean factor score $avg(E[z|\mathbf{u}, g])$ as stated in hypothesis 2.

Note that in general, there exist several empirical methods to evaluate the effect of different number of mixture components and latent dimensions based on different criteria, such as eigen scree-plot and likelihood cross-validation as shown in Section 4.1. As a matter of fact, in our case, different numbers change the overall likelihood scale for both models but not their relative comparison results. Since our main purpose here is the verification of the hypotheses we are interest in, we have kept a reasonably small number, which is convenient to manage without loss of generality.

4.3 Evaluation of Segment Quality

To asses the quality of the discovered segments we use the metric *Weighted Average Visit Percentage* (WAVP) [8]. WAVP allows us to evaluate each segment based on the likelihood that a user who visits any page in the centroid profile, will visit the rest of the pages in that profile during the same session. Specifically, let T be the set of transactions in the evaluation set, and for a certain segment profile in the form of a page vector $\boldsymbol{v}^g$, let T^g denote a subset of T whose sessions

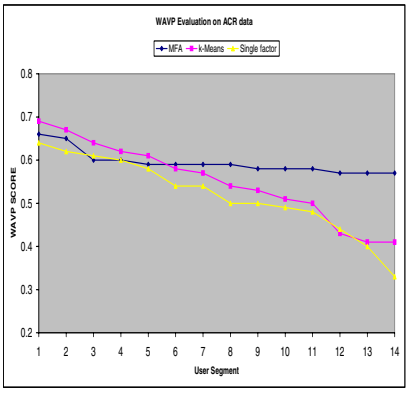

Fig. 3. WAVP evaluation on ACR data

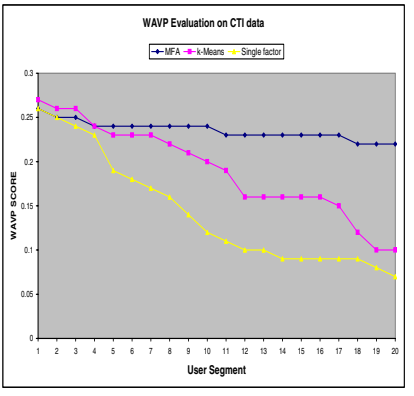

Fig. 4. WAVP evaluation on CTI data

should contain at least one page from the profile. The weighted average similarity to the profile v^g over all sessions in T^g is computed, and this average is then divided by the total weight of page in the profile:

$$WAVP\,(\boldsymbol{v}^g, T) = \frac{\sum_{t \in T^g} \boldsymbol{t} \cdot \boldsymbol{v}^g / |T^g|}{\sum_d weight\,(d, \boldsymbol{v}^g)},$$

where $weight\,(d, v^g)$ is the weight of a page d in this profile $\boldsymbol{v}^g$. The higher WAVP value the better the profile is, in the sense that the corresponding segment is more representative about the similar user navigational activities.

Figures 3 and 4 show the WAVP evaluation for the two data sets, comparing the MFA-based approach to FA and to standard segmentation approach using k-means clustering. All segments are ranked in descending order of WAVP. The results show that the MFA-based user models have consistently higher WAVP scores in general. Also, since MFA-based segments will generally capture more complex patterns capturing multiple factors influencing user's online behavior, the variation of WAVP scores across all MFA-based segments are significantly smaller then those of k-Means and single factor models.

5 Conclusions

The generative modeling based on FA and MFA for Web users' navigation behavior is intuitively reasonable. We can assume that users' navigation data are generated according to some distribution that is conditioned on users' hidden preferences, which can be modeled as hidden variables in latent variable models. In this paper, We have introduced an MFA-based usage mining approaches that can discover both quantitative relationship between users' manifest observations and latent factors, as well as mixture components representing user segments. Our experimental results show that our approach can successfully discover heterogeneous user segments and characterize these segments with respect of their

common preferences. The aggregate representation of Web user segments, combining both of user's navigation data and the user-component memberships, can be used for explorative analysis purposes or for dynamically predicting a new user's navigational interests and recommending relevant pages or products accordingly.

References

1. 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.
2. B.J. Frey, A. Colmenarez, and T.S. Huang. Mixtures of local linear subspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Los Alamitos, June 1998.
3. Z. Ghahramani and G. Hinton. The EM algorithm for mixture of factor analyzers. Technical report CRG-TR-96-1, University of Toronto, 1996.
4. W. Lin, S. A. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6:83–105, 2002.
5. G. McLachlan and D. Peel. Mixtures of factor analyzers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, USA, 2000.
6. B. Mobasher. Web usage mining and personalization. In Munindar P. Singh, editor, *In Practical Handbook of Internet Computing*. CRC Press, 2005.
7. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd ACM Workshop on Web Information and Data Management (WIDM01)*, Atlanta, Georgia, November 2001.
8. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6:61–82, 2002.
9. L.K. Saul and M.G. Rahim. Modeling acoustic correlations by factor analysis. In M. I. Jordan and M. S. Kearn sand S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 749–756. MIT Press, 1998.
10. J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.
11. M. Wedel and W. Kamakura. *Market Segmentation: Conceptual and Methodological Foundations*. Springer, 1999.
12. Y. Zhou, X. Jin, and B. Mobasher. A recommendation model based on latent principal factors in web navigation data. In *Proceedings of the 3rd International Workshop on Web Dynamics at WWW2004*, New York, May 2004.