Available online at www.sciencedirect.com

ScienceDirect

Journal of Retailing and Consumer Services 14 (2007) 359–368

JOURNAL OF RETAILING AND CONSUMER SERVICES

www.elsevier.com/locate/jretconser

Latent class modeling of website users' search patterns: Implications for online market segmentation

José G. Dias[a,*], Jeroen K. Vermunt[b]

[a]Department of Quantitative Methods and UNIDE, ISCTE, Higher Institute of Social Sciences and Business Studies, Edifício ISCTE, Av. das Forças Armadas, 1649–026 Lisboa, Portugal
[b]Department of Methodology and Statistics, Tilburg University, P.O. Box 90153 NL-5000 LE Tilburg, The Netherlands

Abstract

Appropriate modeling of web use patterns may yield very relevant marketing and retailing information. We propose using a model-based clustering approach for market segmentation based on website users' search patterns. We not only provide a detailed discussion of technical issues such as the problem of the selection of the number of segments, but also a very interesting empirical illustration of the potentials of the proposed approach.

Keywords: Web usage mining; Market segmentation; Latent class model

1. Introduction

In recent years web usage mining has become a very important topic of research in the field of data mining. Web usage mining involves using data mining techniques in the discovery of web navigation patterns from web log data, also referred to as click stream data (Hand et al., 2001; Spiliopoulou and Pohle, 2001; Huang et al., 2006). The idea is that the analysis of the sequence of web pages requested by users within a particular web site may provide a better understanding and prediction of users' behavior, and may thus be used to improve the design of the web site concerned. For example, web mining of online stores may yield information on the effectiveness of marketing and web merchandizing efforts, such as how the consumers start the search, which products they see, and which products they buy (Shahabi et al., 1997; Lee et al., 2001). It has, however, been shown that traditional data mining algorithms may not be suited for the discovery of web usage patterns (Spiliopoulou and Pohle, 2001).

An approach that has been used extensively for web mining views a particular web user's navigation pattern on a web site as a Markovian process (Sarukkai, 2000; Dongshan and Junyi, 2002). A Markov model is built for predicting the next web page requested by a web user. A limitation of simple Markov models is that they do not take into account differences in user preferences.

In this paper we propose capturing unobserved heterogeneity among web users using a model with a discrete latent variable. This latent variable allows the segmentation of the data set into clusters and, once the model is learned, web users can be assigned into clusters. More specifically, we use a model known as finite mixture of Markov chains, which has been applied as a model-based clustering tool for classifying web users into different categories (Cadez et al., 2003; Sen and Hansen, 2003). This model has been referred to as mixed Markov model in applied literature (Poulsen, 1990; van de Pol and Langeheine, 1990). Despite of the widespread application of finite mixture models, the number of latent segments to retain is still a very important topic of research. Whereas most researchers use AIC and BIC for determining the dimension of these models, there is evidence that the new AIC3 measure is more appropriate for discrete data (Andrews and Currim, 2003).

*Corresponding author. Tel.: +351 217 903 228; fax: +351 217 903 004.
E-mail addresses: jose.dias@iscte.pt (J.G. Dias), J.K.Vermunt@uvt.nl (J.K. Vermunt).

0969-6989/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jretconser.2007.02.007

The goal of this paper is fifth fold. First, it brings together web mining and market segmentation issues in retailing research. Second, we discuss the modeling of web usage patterns by finite mixtures. Third, we discuss the setting of the number of segments and provide a Monte Carlo (MC) study. Fourth, we introduce the Markov map that allows a fast understanding of the dynamics within each segment. Finally, an empirical application shows the relevance of the connection between web mining and marketing segmentation developed here.

Section 2 gives a short review of market segmentation issues in web mining applications. Section 3 presents model specification and estimation of the finite mixture model for sequential data. It also discusses the selection of the number of latent segments using information criteria. Section 4 presents the results from the MC study that was performed to gain more insight in the performance of different model selection criteria. Section 5 illustrates the implications for web usage mining of the theoretical results with a real data set. The paper concludes with a summary of main findings, implications, and suggestions for further research.

2. Market segmentation issues in web mining research

Market segmentation has become a key concept in marketing theory and practice (Wind, 1978; Wedel and Kamakura, 2000). Smith (1956) defined it as: "market segmentation consists of viewing a heterogeneous market as a number of smaller homogeneous markets in response to differing product preferences, (…) attributable to the desires of consumers or users for more precise satisfaction of their varying wants" (p. 6). Viewing a heterogeneous population as being composed of homogeneous subgroups or segments, each of which responds differently to the marketing mix has been shown to greatly increase marketing efficiency (Wedel and Kamakura, 2000).

Segmentation studies to identify those homogeneous groups have two major components: the information used as input, called the 'bases' of segmentation, and the methods used to identify segments/subpopulations based on the input data (Wind, 1978; Wedel and Kamakura, 2000).

Traditionally, segmentation bases have been classified into four categories: demographic variables (e.g., age, sex, household size), geographic variables (e.g., ZIP code, region), psychographic variables (e.g., attitudes, values, lifestyles), and behavioral variables (e.g., frequency of use, usage level). The most effective segmentation strategy is that which best captures differences in the behavior of target subpopulations, for instance, behavior status or behavioral dynamics.

Segmentation methods can be classified into (Wedel and Kamakura, 2000): (i) *a priori* approaches, when the type and number of subpopulations are defined in advance based on specific criteria, usually demographic variables; and (ii) *post hoc* approaches, when the type and number of subpopulations emerge from the segmentation procedure applied to the data after they have been collected. *Post hoc* approaches to segmentation, and in particular those based on finite mixture models have received much attention in the marketing literature (Wedel and Kamakura, 2000). Finite mixture models have been shown to outperform traditional *post hoc* approaches involving cluster analysis (Vriens et al., 1996).

It is therefore not surprising that also in web mining a substantial effort has been put in an attempt to discover groups of users exhibiting similar browsing patterns (Vakali et al., 2004). Whereas the purpose of user clustering is to establish groups of users that present similar browsing patterns, sometimes the aim may be to discover groups of pages with a similar content (Shahabi et al., 1997; Petridou et al., 2006). Poblete and Baeza-Yates (2006) adopted such a supply-oriented viewpoint based on the clustering of pages of web sites. Various examples exist of studies adopting a consumer-oriented viewpoint—which is also our focus—and that involve clustering users. Petridou et al. (2006) proposed clustering web users using a K-means algorithm based on the KL-divergence which measures the "distance" between individual data distributions. A similar approach is adopted by Yang and Padmanabhan (2005), who looked at the number of times a given user visited a given webpage. Smith and Ng (2003) suggested using self-organizing maps (SOMs) of user navigation patterns. However, none of these approaches accounts for the sequential structure of browsing data. This means that consecutive states in a sequence are, in fact, treated as independent observations conditional on cluster membership, an assumption that is rather unrealistic.

Because finite mixtures are probabilistic or stochastic models, it is possible to parameterize the model in ways that respect the structure of data (e.g., taking into account serial dependence) and test for response parameters and the number of subpopulations with appropriate statistical methods (McLachlan and Peel, 2000). From an unsupervised learning perspective, Markov models with latent variables have been an important paradigm for modeling sequential data (Saul and Jordan, 1998; Cadez et al., 2003; Pallis et al., 2005).

3. The latent segment Markov chain model

This section introduces a flexible model for web users' search patterns that accommodates for both heterogeneity and serial dependencies. Consider a sample of n web users. A web user will be denoted by i ($i = 1, \ldots, n$). Each web user is characterized by a sequence of states $\mathbf{x}_i$. Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote a sample of size n. Note that the state of the website user is, in fact, the category to which the current web page belongs to (frontpage, new, etc.). A next state is generated by moving to another new page, where the next state will be the same as the previous if the new pages concerned belong to the same category. A sequence is formed by a series of visited pages.

3.1. Incorporating sequential dependency

Let X_{it} be the random variable denoting the state of the web user i at position t, the category of the tth webpage, and x_{it} a particular realization. We will assume discrete time from 0 to T_i ($t = 0, 1, \ldots, T_i$). Note that the length of the sequence may differ among web users. Thus, the vectors $\mathbf{X}_i$ and $\mathbf{x}_i$ denote the consecutive requested pages—respectively X_{it} and x_{it}—with $t = 0, \ldots, T_i$. The probability density $P(\mathbf{X}_i = \mathbf{x}_i) = P(X_{i0} = x_{i0}, X_{i1} = x_{i1}, \ldots, X_{iT_i} = x_{iT_i})$ can be extremely difficult to characterize, due to its possibly huge dimension $(T_i + 1)$. A common procedure to simplify $P(\mathbf{X}_i = \mathbf{x}_i)$ is by assuming the Markov property stating that the occurrence of event $X_t = x_t$ depends only upon the previous state $X_{t-1} = x_{t-1}$; that is, conditional on X_{t-1}, X_t is independent of the states at the other time points. From the Markov property, it follows that

$$P(\mathbf{X} = \mathbf{x}_i) = P(X_{i0} = x_{i0}) \prod_{t=1}^{T_i} P(X_{it} = x_{it} | X_{i,t-1} = x_{i,t-1}),$$

where $P(X_{i0} = x_{i0})$ is the initial distribution and $P(X_{it} = x_{it} | X_{i,t-1} = x_{i,t-1})$ is the probability that web user i is in state x_{it} at t, given that he is in state $x_{i,t-1}$ at time $t-1$ (for an introduction to Markov chains, see Ross, 2000). A first-order Markov chain is specified by its transition probabilities and initial distribution. Hereafter, we denote the initial and the transition probabilities as $\lambda_j = P(X_{i0} = j)$ and $a_{jk} = P(X_t = k | X_{t-1} = j)$, respectively. Note that we assume that transition probabilities are time homogeneous, which means that our model is a stationary first-order Markov model.

3.2. Incorporating unobserved heterogeneity

The LSMC—latent segment Markov chain model—assumes discrete heterogeneity. Web users are clustered into S segments, each denoted by s ($s = 1, \ldots, S$). The latent segments, including its number, are not known *a priori*. Thus, in advance one does not know how the sample will be partitioned into clusters. The latent segment that web user i belongs to is denoted by the latent discrete variable $Z_i \in \{1, 2, \ldots, S\}$. Let $\mathbf{z} = (z_1, \ldots, z_n)$. Because $\mathbf{z}$ is not observed, the inference problem is to estimate the parameters of the model, say φ, using only information on $\mathbf{x}$. More precisely, the estimation procedure has to be based on the marginal distribution of $\mathbf{x}_i$ which is obtained as follows:

$$P(\mathbf{X}_i = \mathbf{x}_i; \varphi) = \sum_{s=1}^{S} \pi_s P(\mathbf{X}_i = \mathbf{x}_i | Z_i = s).$$

This equation defines a finite mixture model with S latent segments (McLachlan and Peel, 2000). The latent segment proportions, $\pi_s = P(Z_i = s)$, correspond to the *a priori* probability that web user i belongs to the segment s, and gives the segment relative size. Moreover, π_s satisfies $\pi_s > 0$ and $\sum_{s=1}^{S} \pi_s = 1$.

Within each latent segment s, observation $\mathbf{x}_i$ is characterized by $P(\mathbf{X}_i = \mathbf{x}_i | Z_i = s) = P(\mathbf{X}_i = \mathbf{x}_i | Z_i = s; \theta_s)$ which implies that all individuals in segment s have the same probability distribution defined by the segment-specific parameters θ_s. The parameters of the LSMC model are $\varphi = (\pi_1, \ldots, \pi_{S-1}, \theta_1, \ldots, \theta_S)$. The θ_s includes the transition and initial probabilities $a_{sjk} = P(X_{it} = k | X_{i,t-1} = j, Z_i = s)$ and $\lambda_{sj} = P(X_{i0} = j | Z_i = s)$, respectively. The independent parameters of the LSMC model are $S - 1$ prior probabilities, $S(K-1)$ initial probabilities, and $SK(K-1)$ transition probabilities. Thus, the total number of independent parameters is $SK^2 - 1$.

It is important to note that a probability distribution corresponding to a finite mixture of Markov chains ($S > 1$) cannot be described by a (single segment) Markov chain. This shows that the finite mixture model of Markov chains enables the modeling of much more complex patterns than the standard Markov model (Cadez et al., 2003; Dias and Willekens, 2005). This is illustrated by Fig. 1 which shows that X_t and $X_{t'}$ are assumed to be mutually independent conditionally on X_{t-1}, X_{t+1} and Z. However, in this mixture model, unconditionally on Z, X_t and $X_{t'}$ do not need to be independent given X_{t-1} and X_{t+1}; that is, the standard Markov assumptions does not need to hold.

3.3. LSMC model estimation

The log-likelihood function for φ, given that $\mathbf{x}_i$ are independent observations, is $\ell_S(\varphi; \mathbf{x}) = \sum_{i=1}^{n} \log P(\mathbf{X}_i = \mathbf{x}_i; \varphi)$, and the maximum likelihood estimator (MLE) is $\hat{\varphi} = \arg\max_{\varphi} \ell_S(\varphi; \mathbf{x})$. The EM algorithm is a very attractive option for LSMC model estimation (Dempster et al., 1977; McLachlan and Krishnan, 1997). For estimating this model by the EM algorithm, we refer to Dias and Willekens (2005). For $S = 1$ we have the homogeneous or aggregate Markov chain model. The programs for this study were written in MATLAB (MathWorks, 2002). Also commercial mixture modeling software, such as Latent GOLD (Vermunt and Magidson, 2005), can be used for the estimation of the LSMC model, but may require a specific construction of the input data set.

3.4. Decision on the number of latent segments

An important theoretical and practical issue in market segmentation is the setting of the number of market segments (S). The standard approach to the selection of the best among different nested models is using the likelihood

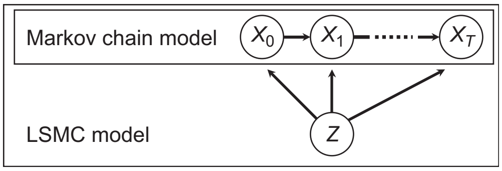

Fig. 1. Extension from the Markov chain model to the LSMC model.

ratio test, which under certain regularity conditions has a simple asymptotic theory (Wilks, 1938). However, this approach cannot be used in the context of latent class/finite mixture models because these regularity conditions do not hold. This is one of the reason why the use of information criteria based on the principle of parsimony are so popular in this field. The underlying idea is that all other things being the same (log-likelihood), we should prefer the simplest model (with fewer parameters). More specifically, the selected number of latent segments is the minimizer of $C_S = -2\ell_S(\hat{\varphi}; \mathbf{x}) + dN_S$, where $\ell_S(\hat{\varphi}; \mathbf{x})$ and N_S are the log-likelihood and the number of free parameters of the model, respectively. For different values of d, we obtain the AIC (Akaike, 1974; $d = 2$), BIC (Schwarz, 1978; $d = \log n$), CAIC (Bozdogan, 1987; $d = \log n + 1$), and AIC3 (Bozdogan, 1993; $d = 3$) measures.

Apart from the four information criteria reported above, we also investigated a different definition of the BIC and CAIC. Some researchers have considered as *sample size* the sum of the number of repeated measurements from all observations (Ramaswamy et al., 1993; DeSarbo et al., 2004), in which case the penalization becomes a function of $n(T+1)$ instead of n (for sequences of the same length).

4. MC study

As we have seen the number of market segments has important implications for marketing management. Despite extensive study of the performance of information criteria for model selection in finite mixtures, little is known about the performance of these criteria for finite mixtures of Markov chains.

4.1. Experimental design

A Monte Carlo study is performed to assess the ability of the different information criteria to retrieve the true model. The experimental design controls the length of the sequences $(T+1)$, the number of webpage types (K), the sample size (n), and the balance of latent segment sizes. The number of variables $(T+1)$ was set at levels 10, 20, and 40; and the number of states at levels 2 and 4. The sample size (n) assumes the levels: 300, 600, and 1200. The number of latent segments in the Monte Carlo study is set to two $(S = 2)$, and models with one, two, and three components are estimated. The latent segment size can be equal (level 1) or unequal (level 2): $\pi = (0.5, 0.5)$ and $\pi = (0.4, 0.6)$, respectively.

The (true) parameter values are shown in Table 1. These *ad hoc* values try to cover different situations in empirical data sets. In particular, there is an attempt to include persistent patterns usually observed in web mining applications with high retention probabilities.

To avoid local optima, for each number of latent segments (2 and 3) the EM algorithm was repeated 5 times with random starting centers, and the best solution (maximum likelihood value out of those 5 runs) and model selection results were kept. The EM algorithm ran until the difference between log-likelihoods being smaller than 10^{-4}.

Table 1
The (true) parameter values for the Monte Carlo study

Parameters	$K = 2$		$K = 4$			
	$k = 1$	$k = 2$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
λ_{1k}	0.60	0.40	0.30	0.20	0.30	0.20
λ_{2k}	0.40	0.60	0.20	0.30	0.20	0.30
a_{11k}	0.75	0.25	0.45	0.15	0.25	0.15
a_{12k}	0.33	0.67	0.20	0.40	0.30	0.10
a_{13k}	—	—	0.40	0.40	0.10	0.10
a_{14k}	—	—	0.30	0.30	0.20	0.20
a_{21k}	0.25	0.75	0.15	0.45	0.15	0.25
a_{22k}	0.67	0.33	0.40	0.20	0.10	0.30
a_{23k}	—	—	0.10	0.10	0.40	0.40
a_{24k}	—	—	0.20	0.20	0.30	0.30

This MC study sets a $2^2 \times 3^2$ factorial design with 36 cells. Special care needs to be taken before arriving at conclusions based on MC results. In this study, we performed 25 replications within each cell to obtain the frequency distribution of selecting the true model, resulting in a total of 900 data sets. The main performance measure used is the frequency with which an information criterion picks the correct model. For each generated data set, the criteria are classified as *underfitting*, *fitting*, or *overfitting*, depending on whether the estimated S is smaller than, equal to, or larger than the true S.

4.2. Results

In agreement with previous research on the number of market segments, our key finding is the overall remarkable performance of AIC3 (Andrews and Currim, 2003). AIC3 outperforms more widespread information criteria such as AIC and BIC for the LSMC model. It identifies the true model 83.2% of the times (Table 2). The main problem associated with AIC is that it tends to overfit the data (20.9%). The AIC3 presents minor overfitting (0.7%). For CAIC and BIC the penalization $n(T+1)$ reduces their performance and it is not considered hereafter.

A second objective of the study was the comparison of the performance of the information criteria across design factors. It turns out that increasing the sample size always improves the performance of the information criteria by reducing underfitting. An exception is AIC for which increasing the sample size tends to increase overfitting. Increasing the number of measurements $(T+1)$ improves the performance of the information criteria and reduces the underfitting. Increasing the state space (K) reduces the underfitting, and thus improves the performance of the information criteria. The balance of latent segment sizes has a dramatic effect on the performance of the information criteria: their performance is much worse with unequal class sizes than with equal class sizes, which shows up in the form of an increased underfitting for the former setting.

Table 2
Results of the Monte Carlo study

Factors		Criteria					
		AIC	AIC3	CAIC		BIC	
				n	$n(T+1)$	n	$n(T+1)$
Sample size (n)							
300	Underfit	0.170	0.237	0.340	0.383	0.327	0.360
	Fit	0.680	0.760	0.660	0.617	0.673	0.640
	Overfit	0.150	0.003	0.000	0.000	0.000	0.000
600	Underfit	0.126	0.143	0.310	0.327	0.263	0.327
	Fit	0.691	0.850	0.690	0.673	0.737	0.673
	Overfit	0.183	0.007	0.000	0.000	0.000	0.000
1200	Underfit	0.077	0.103	0.183	0.273	0.167	0.237
	Fit	0.727	0.887	0.817	0.727	0.833	0.763
	Overfit	0.196	0.010	0.000	0.000	0.000	0.000
Number of variables ($T+1$)							
10	Underfit	0.193	0.257	0.423	0.467	0.407	0.447
	Fit	0.643	0.737	0.577	0.533	0.593	0.553
	Overfit	0.164	0.006	0.000	0.000	0.000	0.000
20	Underfit	0.117	0.177	0.257	0.300	0.250	0.277
	Fit	0.663	0.813	0.743	0.700	0.750	0.723
	Overfit	0.220	0.010	0.000	0.000	0.000	0.000
40	Underfit	0.023	0.050	0.153	0.217	0.100	0.200
	Fit	0.733	0.947	0.847	0.783	0.900	0.800
	Overfit	0.244	0.003	0.000	0.000	0.000	0.000
Number of states (K)							
2	Underfit	0.220	0.302	0.436	0.478	0.400	0.467
	Fit	0.624	0.687	0.564	0.522	0.600	0.533
	Overfit	0.156	0.011	0.000	0.000	0.000	0.000
4	Underfit	0.002	0.020	0.120	0.178	0.104	0.149
	Fit	0.736	0.978	0.880	0.822	0.896	0.851
	Overfit	0.262	0.002	0.000	0.000	0.000	0.000
Proportions							
1	Underfit	0.000	0.000	0.000	0.000	0.000	0.000
	Fit	0.709	0.991	1.000	1.000	1.000	1.000
	Overfit	0.291	0.009	0.000	0.000	0.000	0.000
2	Underfit	0.222	0.322	0.556	0.656	0.504	0.616
	Fit	0.651	0.673	0.444	0.344	0.496	0.384
	Overfit	0.127	0.005	0.000	0.000	0.000	0.000
Overall							
	Underfit	0.111	0.161	0.278	0.328	0.252	0.308
	Fit	0.680	0.832	0.722	0.672	0.748	0.692
	Overfit	0.209	0.007	0.000	0.000	0.000	0.000

5. Web mining application

5.1. Data

The data set used in this research contains information on the search sequences of visitors on `msnbc` (the *msnbc.com anonymous web data set* is available at `kdd.ics.uci.edu/databases/msnbc/msnbc.data.html`). It describes the page visits on `msnbc.com` on September 28, 1999. Each sequence in the data set corresponds to page views of a user during a 24 h period. The original number of users is 989 818 and each event in the sequence is classified into the following categories: (1) `frontpage`, (2) `news`, (3) `tech`, (4) `local`, (5) `opinion`, (6) `on-air`, (7) `misc`, (8) `weather`, (9) `health`, (10) `living`, (11) `business`, (12) `sports`, (13) `summary`, (14) `bbs` (bulletin board service), (15) `travel`, (16) `msn-news`, and (17) `msn-sports`. This data set has been used by others (Cadez et al., 2003). We considered in our analysis a random sample of 1% of the users with a sequence of length at least 2. The sample size is $n = 6244$.

Under a homogeneous population assumption (single segment) the Markov chain model (the same as the 1-LSMC model) of web users' patterns can be summarized by a single initial distribution (Table 4, column Aggregate) and a single set of transition probabilities. Fig. 2 depicts a novel graphical display for the matrix of transition probabilities that we call a Markov map. For the minimum and maximum values of the transitions probabilities (0 and 1), we use white and black, respectively. Values in between with a gray color which is obtained by a linear grading of colors between white and black. Note that the origin states are in the rows and the destination states in the column, which means that the row totals are equal to 1. From the analysis of this single Markov chain, we observe that 34.7% of the users start their search from `frontpage`, followed by `on-air` (11.1%), `weather` (9.1%), `msn-news` (7.9%), and `msn-sports` (6.7%). The transitions to the same state are very persistent in the data set (almost absorbing states). Indeed, most of the diagonal probabilities are close to one, which means that from an aggregate viewpoint the web user tends to stay at the same state. The Markov map shows, for example, that there is a tendency to return to `frontpage` for most of the states, to move out of `summary`, and a certain interaction between `on-air` and `mis`, `msn-sports` and `sports`, and

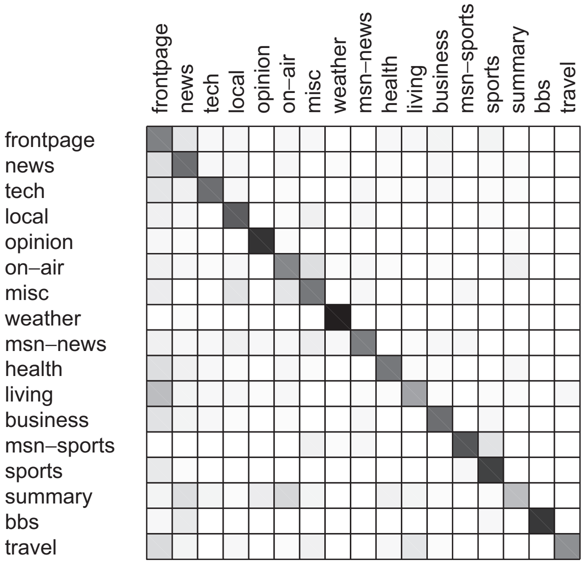

Fig. 2. Markov map for the aggregate model (1-LSMC).

`frontpage` and `news`. This description of the dynamics of web users' patterns is not very informative, because all web users are assumed to follow exactly the same sequential pattern (taste homogeneity). Apart from that, the Markov property under the homogeneous population might be problematic.

5.2. Selecting the number of latent segments

Instead of representing web users' patterns by a single Markov chain, heterogeneity is hereafter introduced by allowing more than one segment, i.e., each web user belongs to one of S groups, each of which contains persons with similar browsing pattern. We estimated LSMC models with 1–8 segments using 20 different starting values to avoid local maxima. Based on BIC and CAIC values presented in Table 3 (best model in italics), one would conclude that two segments have to be included in the model ($S = 2$). According to AIC3, no less than six segments need to be used to describe the data in a satisfactory way.

Table 3
Model selection criteria

S	$\ell_S(\varphi; \mathbf{x})$	Information criteria			
		BIC	AIC	AIC3	CAIC
1	−63838.6	130194.1	128253.2	128541.2	130482.1
2	−61607.3	*128257.3*	124368.7	124945.7	*128834.3*
3	−60370.5	128309.2	122472.9	123338.9	129175.2
4	−59573.0	129239.9	121455.9	122610.9	130394.9
5	−59072.5	130764.7	121033.0	122477.0	132208.7
6	−58543.5	132232.4	120553.1	*122286.1*	133965.4
7	−58235.5	134142.0	*120515.0*	122537.0	136164.0
8	−57978.4	136153.5	120578.8	122889.8	138464.5

5.3. Results

Table 4 and Fig. 3 summarize the LSMC with two latent segments (2-LSMC). The prior probability of each segment (π_s) indicates segment sizes (Table 4). Segment I, the largest (57.6%), is still very heterogeneous with web users starting their browsing mainly from `on-air` (15.7% of the web users in this segment start their sequence in this state), `frontpage` (14.0%) and `msn-news` (11.7%). Moreover, this segment has a very stable pattern of browsing (Fig. 3) almost absorbing for most of the states. However, the web users starting from `on-air` and `frontpage`, in particular the second one, tend to move to other states. Segment II (42.4% of the sample) is rather stable. Indeed, most of them start their search from `frontpage` (62.9%) or `weather` (15.3%) states and tend to stay in these states. On the other hand, even users starting from other states tend to move to `frontpage` (Fig. 3).

For six segments or patterns of web browsing (supported by AIC3), the segment sizes range from 12.5% to 21.3% of the sample. Segment I (16.6% of the sample) represents a generalist segment starting mainly from `frontpage` (55.8%) and `news` (9.9%). From Fig. 4 we observe a strong tendency to return to `frontpage` from most of the states. Web users in this segment seem to wish to be kept informed on general issues, in particular news and sports. Web users in segment II, the largest segment (21.3% of the sample), are specially concerned with `weather` (25.2%)

Table 4
Initial distribution and proportions

	Aggregate	$S = 2$		$S = 6$					
		1	2	1	2	3	4	5	6
Initial distribution (λ_{sj})									
`frontpage`	0.347	0.140	0.629	0.558	0.259	0.108	0.920	0.038	0.254
`news`	0.060	0.093	0.016	0.099	0.021	0.044	0.003	0.018	0.231
`tech`	0.039	0.059	0.011	0.043	0.074	0.050	0.009	0.024	0.018
`local`	0.053	0.083	0.012	0.023	0.081	0.062	0.000	0.078	0.062
`opinion`	0.006	0.007	0.004	0.005	0.017	0.006	0.000	0.002	0.000
`on-air`	0.111	0.157	0.049	0.052	0.068	0.100	0.012	0.285	0.155
`misc`	0.006	0.005	0.007	0.008	0.003	0.018	0.001	0.004	0.000
`weather`	0.091	0.046	0.153	0.020	0.252	0.078	0.000	0.087	0.049
`msn-news`	0.079	0.117	0.029	0.027	0.011	0.268	0.013	0.106	0.068
`health`	0.015	0.022	0.005	0.019	0.002	0.056	0.003	0.010	0.000
`living`	0.011	0.015	0.005	0.009	0.021	0.005	0.003	0.006	0.019
`business`	0.048	0.077	0.008	0.013	0.025	0.109	0.012	0.066	0.072
`msn-sports`	0.067	0.084	0.043	0.046	0.144	0.085	0.008	0.063	0.018
`sports`	0.055	0.085	0.015	0.078	0.010	0.004	0.010	0.197	0.024
`summary`	0.011	0.009	0.012	0.001	0.013	0.002	0.003	0.016	0.030
`bbs`	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.003	0.002
`travel`	0.001	0.001	0.003	0.000	0.000	0.006	0.003	0.000	0.000
Proportions (π_s)	1.000	0.576	0.424	0.166	0.213	0.162	0.156	0.177	0.125

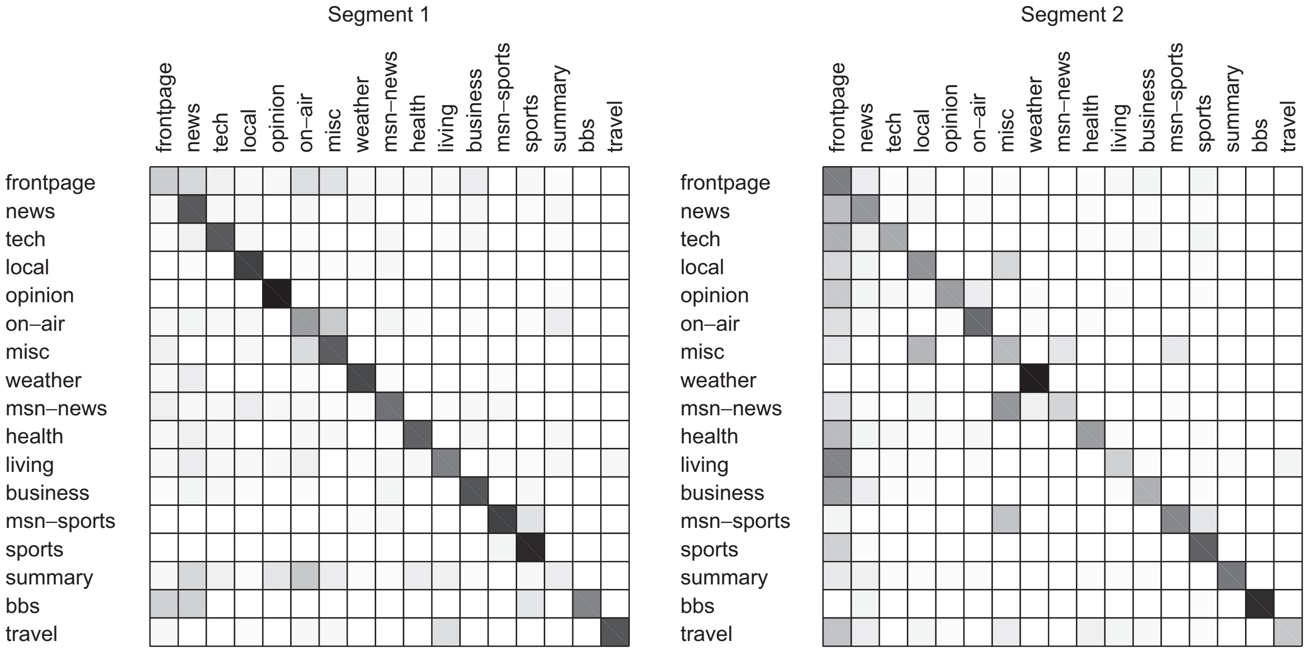

Fig. 3. Markov map for the 2-LSMC model.

and msn-sports (14.4%). We observe in Fig. 4 that web users in this segment tend to be very special in their tastes, given the very high absorbing probabilities of weather and states related to sports (sports, msn-sports, absorbing and high transition probabilities between them). Segment III (16.2% of the sample) represents web users focused on msn-news (26.8%) and business (10.9%). Given the high retention probability of msn-news and business (Fig. 4), this segment is rather stable and mainly concerned with business news. Web users in segment IV (15.6% of the sample) tend to start from frontpage (92.0%). It has a very dynamic browsing pattern (Fig. 4), where frontpage clearly has a pivotal role (very high transition probabilities toward this state from most of the states and low retention probabilities). The web users in this segment are especially concerned with general information with high transition probabilities from and to summary. Web users in segment V, the second largest group (17.7% of the sample), tend to start from on-air (28.5%) and sports (19.7%). The sports state tends to be rather absorbing (Fig. 4) and the states on-air and misc show strong dynamics between them. Finally, segment VI (12.5% of the sample) consists largely of web users concerned with news (23.1%) and on-air (15.5%) and for whom news has a very high retention probability.

Once the segments are identified—and we did not know in advance their number and respective behavior—each segment can be characterized using observable variables. This characterization is, in general, needed to better understand each segment of web users. An interesting next step would be to investigate which persons are in the various latent classes. This can be achieved by a concomitant variable extension of the LSMC model similar to the one proposed for other mixture models (Dayton and MacReady, 1988; Kamakura et al., 1994).

6. Conclusions

In marketing literature, research on online market segmentation has been mostly focused on attitudinal data (Bhatnagar and Ghose, 2004). Modeling and analyzing web usage patterns are helpful for understanding what type of information online market users demand in their interaction with web sites. It is important that web site browsing is intuitive for its users in such a way that they can easily reach the information they search. We introduced a novel statistical model for the analysis of sequences of web searching, which may provide extremely useful retail information for online stores and allow web site developers to understand the browsing behavior of users and may help customize services to them by developing new designs to satisfy future needs. Because the LSMC model accommodates discrete population heterogeneity, it yields a market segmentation structure that can be used for developing segment-specific strategies.

Information criteria such as BIC and AIC have become popular as model selection criteria. We showed by a Monte Carlo study that for the LSMC model the selection of the number of latent segments based on the AIC3 more likely retrieves the true segment structure for web users' search pattern data. Indeed, there is a growing evidence that the Laplace transformation needed in deriving the BIC (Tierney and Kadane, 1982) may perform poorly (suboptimal) for discrete data under a latent segment setting. It

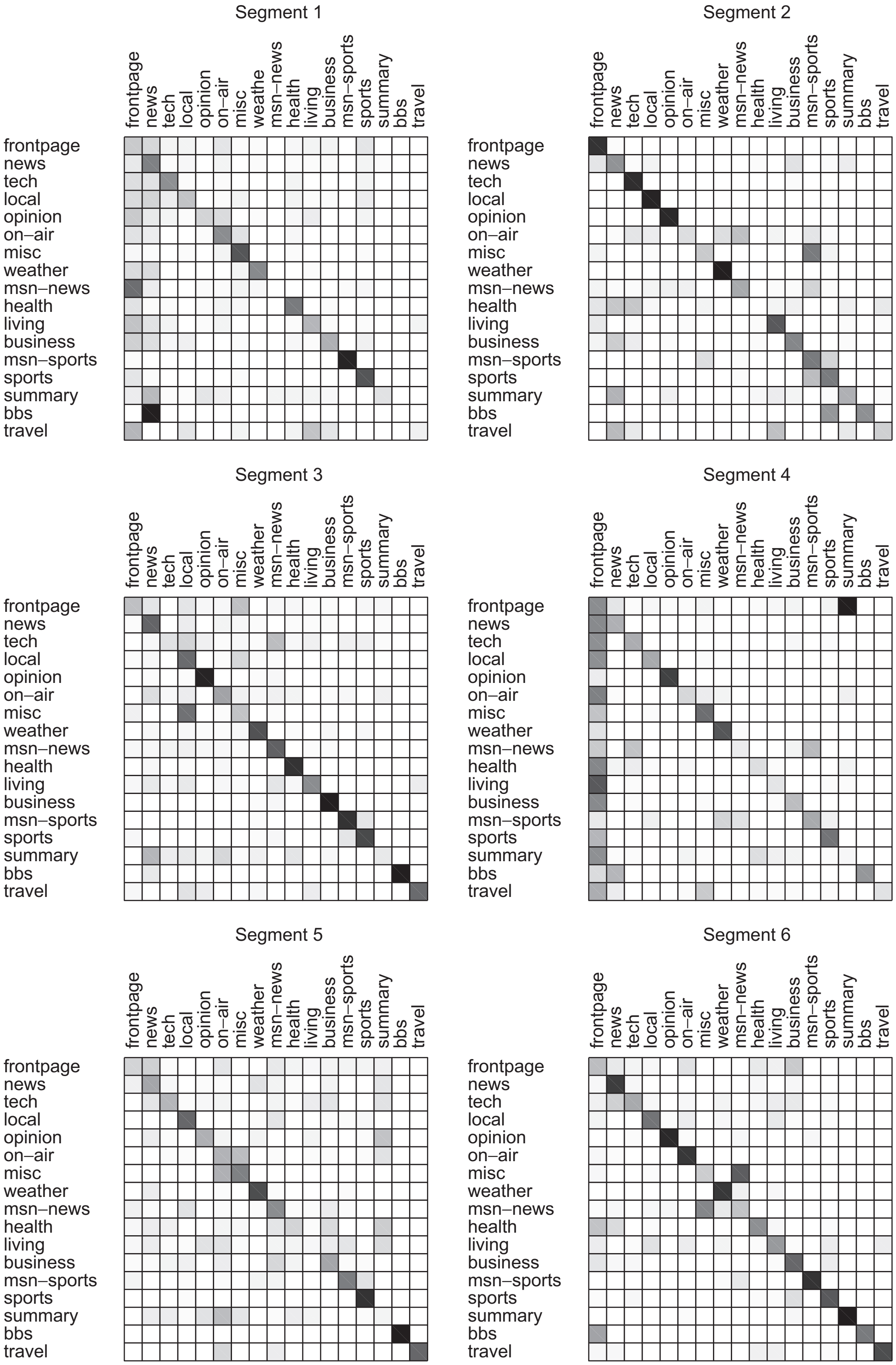

Fig. 4. Markov map for the 6-LSMC model.

was also shown that information criteria which are function of sample size work better with n than with $n(T+1)$ as penalization.

The web mining application illustrated the use of the LSMC model in modeling consumer search patterns on web sites that allows the improvement of the web site design, definition of recommendation systems, and personalizing web sites among other applications (Mobasher et al., 2000). This research also shows the implications of using an inappropriate information criterion in setting the number of market segments. Additionally, we introduce a new graphical tool—Markov map—which summarizes the representation of the dynamics within each latent segment as assumed by the LSMC model. This allows a fast understanding of the observed dynamic patterns between the states.

The LSMC model can be extended in several ways. Future research could be aimed at incorporating characteristics of the web users in the modeling process, which would allow a richer characterization of each segment. This information can be very important in the targeting of consumers in each market segment. Another very important extension would be the incorporation of web design variables as predictors entering in the model for the transition probabilities (Pirolli et al., 2002). That would allow determining the impact of a particular design factor on (segment-specific) transitions.

References

Akaike, H., 1974. A new look at statistical model identification. IEEE Transactions on Automatic Control A-19, 716–723.

Andrews, R.L., Currim, I.S., 2003. A comparison of segment retention criteria for finite mixture logit models. Journal of Marketing Research 40 (2), 235–243.

Bhatnagar, A., Ghose, S., 2004. Segmenting consumers based on the benefits and risks of Internet shopping. Journal of Business Research 57 (12), 1352–1360.

Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52, 345–370.

Bozdogan, H., 1993. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In: Opitz, O., Lausen, B., Klar, R. (Eds.), Information and Classification, Concepts, Methods and Applications. Springer, Berlin, pp. 40–54.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S., 2003. Model-based clustering and visualization of navigation patterns on a web site. Data Mining and Knowledge Discovery 7 (4), 399–424.

Dayton, C.M., MacReady, G.B., 1988. Concomitant-variable latent-class models. Journal of the American Statistical Association 83 (401), 173–178.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1–38.

DeSarbo, W.S., Lehmann, D.R., Hollman, F.G., 2004. Modeling dynamic effects in repeated-measures experiments involving preference/choice: an illustration involving stated preference analysis. Applied Psychological Measurement 28, 186–209.

Dias, J.G., Willekens, F., 2005. Model-based clustering of life histories with an application to contraceptive use dynamics. Mathematical Population Studies 12 (3), 135–157.

Dongshan, X., Junyi, S., 2002. A new Markov model for web access prediction. Computing in Science & Engineering 4 (6), 34–39.

Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. MIT Press, Cambridge, MA.

Huang, Y.M., Kuo, Y.H., Chen, J.N., Jeng, Y.L., 2006. NP-miner: a real-time recommendation algorithm by using web usage mining. Knowledge-based Systems 19 (4), 272–286.

Kamakura, W.A., Wedel, M., Agrawal, J., 1994. Concomitant variable latent class models for conjoint analysis. International Journal of Research in Marketing 11 (5), 451–464.

Lee, J., Podlaseck, M., Schonberg, E., Hoch, R., 2001. Visualization and analysis of clickstream data of online stores for understanding web merchandising. Data Mining and Knowledge Discovery 5 (1–2), 59–84.

MathWorks, 2002. MATLAB 6.5. The MathWorks, Inc., Natick, MA.

McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

Mobasher, B., Cooley, R., Srivastava, J., 2000. Automatic personalization based on web usage mining. Communications of the ACM 43 (8), 142–151.

Pallis, G., Angelis, L., Vakali, A., 2005. Model-based Cluster Analysis for Web Users Sessions. Lecture Notes in Computer Science, vol. 3488. Springer, Berlin, pp. 219–227.

Petridou, S.G., Koutsonikola, V.A., Vakali, A.I., Papadimitriou, G.I., 2006. A Divergence-oriented Approach for Web Users Clustering. Lecture Notes in Computer Science, vol. 3981. Springer, Berlin, pp. 1229–1238.

Pirolli, P.L., Fu, W., Reeder, R., Card, S. K., 2002. A user-tracing architecture for modeling interaction with the World Wide Web. Advanced Visual Interfaces (AVI 2002), May 22–24, Trento, Italy.

Poblete, B., Baeza-Yates, R., 2006. A content and structure website mining model, In: WWW '06: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland. ACM Press, New York, USA, pp. 957–958.

Poulsen, C.S., 1990. Mixed Markov and latent Markov modelling applied to brand choice behaviour. International Journal of Research in Marketing 7 (1), 5–19.

Ramaswamy, V., DeSarbo, W.S., Reibstein, D.J., Robinson, W.T., 1993. An empirical pooling approach for estimating marketing mix elasticities with PIMS data. Marketing Science 12, 103–124.

Ross, S.M., 2000. Introduction to Probability Models, seventh ed. Harcourt, Academic Press, San Diego.

Sarukkai, R.R., 2000. Link prediction and path analysis using Markov chains. Computer Networks—The International Journal of Computer and Telecommunications Networking 33 (1–6), 377–386.

Saul, L.K., Jordan, M.I., 1998. Mixed memory Markov models: decomposing complex stochastic processes as mixtures of simpler ones. Machine Learning 37, 75–87.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Sen, R., Hansen, M.H., 2003. Predicting Web users' next access based on log data. Journal of Computational & Graphical Statistics 12 (1), 143–155.

Shahabi, C., Zarkesh, A.M., Adibi, J., Shah, V. 1997. Knowledge discovery from users Web-page navigation, in: Proceedings of the Seventh International Workshop on Research Issues in Data Engineering (RIDE '97) High Performance Database Management for Large-Scale Applications, IEEE Computer Society, Silver Spring, MD, pp. 20–29.

Smith, K.A., Ng, A., 2003. Web page clustering using a self-organizing map of user navigation patterns. Decision Support Systems 35 (2), 245–256.

Smith, W.R., 1956. Product differentiation and market segmentation as alternative marketing strategies. Journal of Marketing 21 (3), 3–8.

Spiliopoulou, M., Pohle, C., 2001. Data mining for measuring and improving the success of Web sites. Data Mining and Knowledge Discovery 5 (1–2), 85–114.

Tierney, L., Kadane, J., 1982. Accurate approximations for posterior moments and marginal densities. Journal of American Statistical Association 81, 82–86.

Vakali, A., Pokorny, J., Dalamagas, T., 2004. An overview of web data clustering practices. Lectures Notes in Computer Science, vol. 3268. Springer, Berlin, pp. 597–606.

van de Pol, F., Langeheine, R., 1990. Mixed Markov latent class models. Sociological Methodology 20, 213–247.

Vermunt, J.K., Magidson, J., 2005. Latent GOLD 4.0 User's Guide. Statistical Innovations, Inc, Belmont, MA.

Vriens, M., Wedel, M., Wilms, T., 1996. Metric conjoint segmentation methods: a Monte Carlo comparison. Journal of Marketing Research 33 (1), 73–85.

Wedel, M., Kamakura, W.A., 2000. Market Segmentation. Conceptual and Methodological Foundations, International Series in Quantitative Marketing, second ed. Kluwer Academic Publishers, Boston.

Wilks, S.S., 1938. The large sample distribution of the likelihood ratio for testing composite hypotheses. Annals of Mathematical Statistics 9, 60–62.

Wind, Y., 1978. Issues and advances in segmentation theory. Journal of Marketing Research 15 (3), 317–337.

Yang, Y.H., Padmanabhan, B., 2005. GHIC: A hierarchical pattern-based clustering algorithm for grouping Web transactions. IEEE Transactions on Knowledge and Data Engineering 17 (9), 1300–1304.