Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 1348–1367
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Validation and interpretation of Web users' sessions clusters

George Pallis, Lefteris Angelis, Athena Vakali *

Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

Received 25 May 2006; received in revised form 15 October 2006; accepted 15 October 2006
Available online 21 December 2006

Abstract

Understanding users' navigation on the Web is important towards improving the quality of information and the speed of accessing large-scale Web data sources. Clustering of users' navigation into sessions has been proposed in order to identify patterns and similarities which are then managed in the context of Web users oriented applications (searching, e-commerce, etc.). This paper deals with the problem of assessing the quality of user session clusters in order to make inferences regarding the users' navigation behavior. A common model-based clustering algorithm is used to result in clusters of Web users' sessions. These clusters are validated by using a statistical test, which measures the distances of the clusters' distributions to infer their dissimilarity and distinguishing level. Furthermore, a visualization method is proposed in order to interpret the relation between clusters. Using real data sets, we illustrate how the proposed analysis can be applied in popular application scenarios to reveal valuable associations among Web users' navigation sessions.

Keywords: Cluster validation; Web data clustering; Cluster interpretation; Cluster visualization; Web users' sessions mining

1. Introduction

The explosive growth of the Web has drastically changed the way in which information is managed and accessed. The large-scale of Web data[1] sources and the wide availability of services over the Internet have increased the need for effective Web data management techniques and mechanisms. Understanding how users navigate over Web sources is essential both for computing practitioners (e.g. Web sites developers) and researchers (Berendt & Spiliopoulou, 2000). In this context, Web data clustering has been widely used (Baldi, Frasconi, & Smyth, 2003; Banerjee & Ghosh, 2001; Cadez, Heckerman, Meek, Smyth, & White, 2003; Chakrabarti, 2003; Chen & Liu, 2003; Hay, Vanhoof, & Wets, 2001; Pallis, Angelis, & Vakali, 2005; Roussinov & Chen, 2001) for increasing Web information accessibility, understanding users' navigation behavior, improving information retrieval and content delivery on the Web.

* Corresponding author. Tel.: +30 2310998415; fax: +30 2310998419.
E-mail addresses: gpallis@ccf.auth.gr (G. Pallis), lef@csd.auth.gr (L. Angelis), avakali@csd.auth.gr (A. Vakali).

[1] By the term "Web data" we refer to any type of information/file that can be collected and used in the context of the Web (e.g. Web pages, Web access logs, HTML/XML tags etc.).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.10.010

The aim of clustering is to organize information circulated over the Web into groups/collections (of similar objects), in order to facilitate data availability and accessing, and at the same time to meet user preferences. However, clustering the users' navigation patterns is not enough, due to the differences, limitations and diversion of Web-based applications such as e-commerce (to adapt Web sites and to recommend new products to customers' based on their navigation behavior (Baldi et al., 2003; Montgomery, Li, Srinivasan, & Liechty, 2004)) or Web-based information retrieval (to improve real-time and dynamic data accessing (Baldi et al., 2003)).

Considering that there are several users' navigation patterns within groups (each group consists of a huge amount of variable length patterns), the resulted clusters cannot be efficiently evaluated as well as interpreted. Therefore, it is typical to use mechanisms which explore these groups and extract useful conclusions (Cadez, Heckerman, Meek, Smyth, & White, 2003; Fraley & Raftery, 1998).

The goal of this paper is to show that certain statistical techniques, developed for statistical inference of categorical data, are suitable for analyzing the results of a clustering algorithm. The results of such an analysis can be further used to validate and interpret the obtained clusters in order to reveal and explain associations among users' navigation patterns.

The main idea is to improve the access to searching and personalization[2] of the Web sources. These applications take advantages of clustering Web users' navigation patterns, where each navigation pattern reflects the interaction between the users and the Web.

This work originates from the authors' preliminary efforts in (Pallis, Angelis, Vakali, & Pokorny, 2004 & Pallis et al., 2005). The proposed methodology is applied on Web users' navigation patterns by a model-based approach employing:

- *Cluster validation*, i.e. evaluation of the results of a clustering algorithm in a quantitative and objective manner. We propose a quantitative validation procedure, which is based on the statistical chi-square (χ^2) test. Each cluster is represented by a probability distribution and the chi-square metric is used to measure the distances between these distributions and to test their homogeneity. Since the goal of a clustering procedure is to discover groups in the data so that each group is significantly different from all the others, we essentially test the heterogeneity between the clusters in order to assess their successful discrimination.
- *Cluster interpretation*, i.e. understanding and appropriately interpreting the meaning of the derived clusters in the wider context of the underlying application, by using statistical data analysis. Specifically, we propose a visualization approach as a result of the statistical method known as correspondence analysis, for interpreting the clustering results. This analysis is used to facilitate revealing of similar or related features in Web users' navigation behavior and their interaction with the content of Web information sources.

The cluster validation and interpretation method were tested on two real data sets, one from a rather popular and active Web server (msn.com) and one from a typical and low-traffic Web server (Department of Informatics in Aristotle University). Experimental results are encouraging and indicate that the proposed statistical analysis can be used to enhance the existing practices for cluster validation and interpretation.

The rest of the paper is organized as follows: Section 2 provides an overview of the related work in this area and the paper's contribution. Section 3 describes the clustering procedure that is used to group the users' patterns. Section 4 presents the statistical validation technique for model-based clustering approaches. Section 5 presents a visualization method for cluster analysis interpretation. Section 6 provides the experimental results and the application scenarios that could be benefited from the proposed work. Finally Section 7 has the conclusions.

2. Related work and paper's contribution

Earlier research efforts, have mainly been devoted to proposing clustering algorithms and validating clusters, whereas, clustering interpretation has also been considered towards web usage understanding and characterization. A brief related work is highlighted in this section.

[2] Web site personalization can be defined as the process of customizing the content and structure of a Web site to the specific and individual needs of each user taking advantage of the user's navigational behavior (Eirinaki & Vazirgiannis, 2003).

2.1. Clustering algorithms

Existing clustering algorithms for assigning Web users' navigation patterns with common characteristics into the same cluster, may be classified into the typical (Jain, Murty, & Flynn, 1999) following two approaches (also summarized in Table 1):

- *Similarity-based approach* which uses a distance function (similarity measure) to judge whether two patterns should be clustered together. Hierarchical and partitional algorithms have been mostly used under this approach.
- *Model-based approach* which relies on the assumption that objects follow a finite mixture of probability distributions such that each distribution indicates a cluster (each cluster has a data-generating model with its own parameters). In model-based approaches it is critical to learn the parameters for each cluster so that to assign objects to clusters by using a hard assignment policy[3].

Chi-square is commonly used for testing similarities among two different distributions. Moreover, in recent research efforts chi-square is used to define a new similarity measure (Ibrahimov, Sethi, & Dimitrova, 2002), and to classify rows or columns of a contingency table (Govaert & Nadif, in press). Another recent approach (Javed & Bhatti, 2005) is based on chi-square to evaluate a certain hypothesized distribution of data and to decide whether a sub-clustering approach should be followed.

2.2. Validation of web users' sessions clusters

A main challenge with the above clustering algorithms is the difficulty in interpreting and in assessing the quality of the resulted clusters, towards extracting useful inferences for the users' navigation behavior (Chen & Liu, 2003; Halkidi, Batistakis, & Vazirgiannis, 2001; Pallis et al., 2004). Therefore, a clustering approach becomes more valuable if it is further evaluated and validated. For instance, the challenge of a Web personalization system is to provide users with the information of their interests, even when they are not requesting it explicitly. Clustering evaluation may be employed under three different views:

- *External view*: when results of a clustering method are evaluated on the basis of a pre-specified structure on a data set, which reflects a user's intuition about the clustering structure of this data set.
- *Internal view*: clustering results are evaluated in terms of quantities obtained from the data set itself.
- *Relative view*: clustering result is compared with other clustering schemes, by modifying only the parameter values.

Cluster validity approaches based on external and internal criteria rely on statistical hypothesis testing, where the basic idea is to examine whether the points of a data set are randomly structured or not. Such an analysis typically involves a *Null Hypothesis (Ho)* expressed as a statement of random structure of a data set, and this hypothesis is justified by statistical tests, which lead to computationally complex procedures. Several statistical tests have been proposed in the literature for clustering validation, such as Rand Statistitic (R) (Morey & Agresti, 1984) and Cophenetic Correlation Coefficient (CPCC). These statistics are summarized in (Halkidi et al., 2001).

The basic characteristic of the approaches based on internal or external criteria is their high computational demands. On the other hand, the relative approach does not involve statistical tests but evaluates several results originating from different parameter settings and challenge is to choose the best clustering scheme from a set of defined schemes. This choice is commonly done according to a pre-specified criterion, the so-called *cluster validation index*, i.e. a value indicating the quality of a given clustering. Several cluster validation indices have been proposed already where the most indicative are: the Davies–Bouldin index (DB) (Gunter &

[3] In a hard assignment policy, each object is assigned to only one cluster. On the other hand, a soft assignment policy allows degrees of membership in multiple clusters, which means that one object can be assigned to multiple clusters with certain membership values.

Table 1
Approaches in clustering Web users' navigation patterns

Clustering approach	Indicative clustering algorithms
Similarity-based	Sequence alignment method (SAM) Hay et al. (2001); Wang and Zaïane (2002), Generalization-based clustering Fu, Sandhu, and Shih (1999), Weighted longest common subsequences 9 Banerjee and Ghosh (2001), Cube model Huang, Ng, Cheung, and Ching (2001b), Path mining algorithm Shahabi et al. (1997), K-means Chakrabarti (2003)
Model-based	EM algorithm Anderson, Domingos, and Weld (2002); Cadez et al. (2003); Desphpande and Karypis (2001); Sen and Hansen (2003), Self organizing maps clustering Smith and Ng (2003)

Bunke, 2003), the Frobenius norm (Huang, Ng, & Cheung, 2001a), and the other indices overviewed in (Halkidi et al., 2001).

Most of the earlier approaches (e.g. (Gunter & Bunke, 2003; Halkidi et al., 2001; Huang et al., 2001a)) for cluster validation are based on the similarity-based clustering approaches, whereas the model-based approaches have gained ground in the Web community since they can efficiently represent the dynamic "nature" of the Web sources (Fraley & Raftery, 1998; Desphpande & Karypis, 2001; Ypma & Heskes, 2002; Cadez et al., 2003; Pallis et al., 2005). Such model-based approaches capture the users' navigation behavior quite well, by the use of a Markov model, to capture the uncertainties occurring on the Web, which are due to the various large-scale, distributed, decentralized, self-organized, and evolving sources and users navigation patterns. However, further analysis and validation of the model-based clustering schemes is rarely given.

2.3. Interpretation of web users' sessions clusters

Understanding clustering results is not a straightforward process, since different clustering schemes might result in diverse clusters which need further analysis and interpretation. Moreover, clusters role is perceived differently depending on the nature and orientation of the underlying application. This is explained by the fact that in some applications clustering is an initial exploration task (prior to classification which needs clusters i.e. fixed number of classes), whereas in other applications clustering is used to support a decision process (such as in the form of a rule set or a decision tree). Therefore, having an efficient interpretation method is important and often necessary. Several research works in various industrial and academic research communities are focusing on interpreting clusters of users' navigation patterns by:

- Interpreting and analysing users' navigation patterns of online stores (e.g. in (Gomory, Hoch, Lee, Podlaseck, & Schonberg, 1999)), or predicting users' commercial behaviors based on their navigation is proposed (Montgomery et al., 2004);
- Visualizing users grouping (Baldi et al., 2003) by using a mixture model to predict behavior of users, or to interactively visualize Web logs by providing a global view of visitor accesses;
- Visualizing clustering of users' navigation paths in real time by a developed a tool (called INSITE) for knowledge discovery from users Web site (Shahabi, Faisal, Kashani, & Faruque, 2000).

However, according to the authors' knowledge, most of such approaches use empirical methods to interpret the resulted clusters (e.g. a simple visualization of the Web users' patterns in each cluster and make some observations on these data (Cadez et al., 2003)). On the other hand, the users' navigation behavior is a complex procedure and it involves valuable information which is usually hidden (Baldi et al., 2003; Cadez et al., 2003). For instance, several correlations between Web pages may not be observed by using simple visualization schemes. Thus, deeper and more detailed observation is required towards understanding these correlations.

2.4. Paper's contribution

Considering that, in model-based approaches, the clusters are represented by a probabilistic distribution, the proposed validation algorithm originating from (Pallis et al., 2004) is based on a statistical chi-square test.

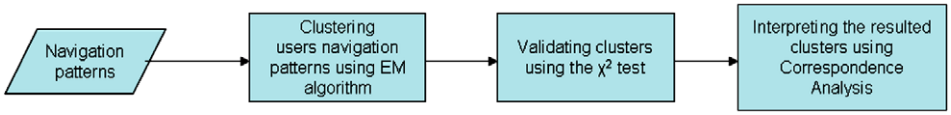


Fig. 1. The proposed procedures sequence.

According to the authors' knowledge, no earlier work has emphasized on using chi-square measure in the process of validating Web users sessions.

The purpose of this paper is to present a combination of probabilistic and statistical methods providing a comprehensive analysis of Web user sessions clusters. The whole approach is presented in the form of a procedure starting from the clustering of the sessions, continuing with the validation of the derived clusters and concluding with the interpretation of the clusters and the utilization of the conclusions for understanding and explaining users' navigation behavior.

First, the users' sessions extracted from a Web site are clustered by the well-known *EM* algorithm, which is a model-based approach able to capture the dynamic evolution of the Web. The optimal number of clusters is determined a priori by using a probabilistic model. Next, the resulted clusters are represented by probability distributions and they are validated using the chi-square test. Finally, the correspondence analysis is used to find relations between clusters and user preferences and therefore to interpret the navigational behavior. Certain applications for which this approach is expected to be beneficial are also highlighted in this paper. The whole methodology can be described by the diagram in the form of a flowchart, given in Fig. 1.

3. Web users' sessions clustering

Clustering of Web sessions under a probabilistic-based approach, involves certain tasks which need to be followed in sequential order, namely session identification, number of clusters identification and clustering algorithm employment.

3.1. Web log files pre-processing and session identification

Users on the Web visit a site and spend arbitrary amount of time at each page between consecutive visits. All the users' traffic is recorded in a Web log file, which is a sequential file with one user access record per line. Web log files provide information about activities performed by a user from the moment the user enters a Web site to the moment the same user leaves it and it typically contains the fields (depicted in Fig. 2):

- domain name (or IP address) of the request;
- name of the user who generated the request;
- date and time of the request;
- method of the request;
- name of the file requested;
- result of the request (success, failure, error, etc.);
- size of the data sent back;
- URL of the referring page;

```
216.239.46.60 - - [04/Jan/2003:14:56:50 +0200] "GET /~lpis/curriculum/C+Unix/Ergastiria/Week-7/filetypes1.txt
HTTP/1.0" 200 86
216.239.46.100 - - [04/Jan/2003:14:57:33 +0200] "GET /~oswinds/top.html HTTP/1.0" 200 869
216.239.46.133 - - [04/Jan/2003:14:58:27 +0200] "GET /~lpis/publications/crc-chapter1.html HTTP/1.0" 304 -
209.237.238.161 - - [04/Jan/2003:14:59:11 +0200] "GET /robots.txt HTTP/1.0" 404 276
```

Fig. 2. A Web server log file.

- identification of the client agent;
- cookie, a string of data generated by an application and exchanged between the client and the server.

Given the Web log file, the goal is to use it as a basic information in order to capture the Web users' navigation trends, typically expressed in the form of Web users' sessions. *A user session is defined as a sequence of requests made by a single user over a certain navigation period and a user may have a single (or multiple) session(s) during a period of time*. A session is a directed list of page accesses performed by an individual user during a visit in a Web site. Several approaches for identifying users' sessions from the Web log files (Banerjee & Ghosh, 2001; Berendt & Spiliopoulou, 2000; Chen, Fu, & Tong, 2003; Hay et al., 2001; He, Göker, & Harper, 2002; Sen & Hansen, 2003; Shahabi, Zarkesh, Adibi, & Shah, 1997) have been proposed in the literature. The most popular session identification methods include:

- *Using a timeout threshold*, in which a user poses a sequence of consecutive requests which are separated by an interval less than a predefined threshold. This session identification suffers from the difficulty to set the time threshold, since different users may have different navigation behaviors, and their time intervals between sessions may significantly vary. In order to define an indicative value for the time threshold, earlier research efforts proposed a time threshold of 25.5 min based on empirical data (Catledge & Pitkow, 1995), whereas in (Goker & He, 2000) used a wide range of values and concluded that a time range of 10–15 min was an optimal session interval threshold. In general, the optimal time threshold clearly depends on the specific context and application. Up to now, the most common choice is to use 30 min as a default time threshold.
- *Considering a reference length* (Chen, Park, & Yu, 1998), i.e. the users' sessions are identified by their maximal forward reference. Each session is defined as the set of documents visited originating from the first document in a request sequence to the final document before a backward reference is made. Here, a backward reference is defined to be a document that has already occurred in the current session. One advantage of the maximal forward reference method is that it does not have any tuneable parameters (e.g., time threshold). However, it has the significant drawback that backward references may not be recorded by the server if caching is enabled at the client site.
- *Dynamically identifying sessions' boundaries* (Huang, Peng, An, & Schuurmans, 2004), based on a statistical n-gram language modelling, to predict the probability of requests' sequences. A session boundary is identified by measuring the change in information (entropy) in the sequence of requests, i.e. when a new object is observed in the sequence, an increase in the entropy of the sequence occurs. Therefore, such an entropy increase serves as an indication for session boundary detection and if the change in entropy is over a specific threshold, then a session boundary is placed before the new object.

In this paper, we use the timeout threshold in order to define the users' sessions and Web log file is automatically processed since manual processing of log files is not feasible (due to their large scale -at least 250,000 records per day are logged in a common Web server). Typically, each Web user can be uniquely identified, by his IP-address, which acts as a unique identifier and at the same time, each of the requested pages has a different unique page id. Then, the data are undergone a certain pre-processing in order to identify Web user sessions. The following steps take place to extract the Web users' sessions from a Web log file:

- *Data cleaning*: We remove all the records which do not include useful information for the users' navigation behavior (such as graphics, javascripts, small pictures of buttons, advertisements etc.) as well as the non-static information (cgi scripts, "?" etc.).
- *Data transformation*: The remaining page ids are categorized into different categories with respect to their content (e.g. a category may be all the pages which refer to the weather in a news site). It should be noticed that the process of grouping the Web pages into categories is a usual practice (Cadez et al., 2003), since it improves the data management and in addition eliminates the complexity of the underlying problem (since the number of page categories is smaller than the number of Web pages in a Web site). In this paper, the individual pages are grouped into semantically similar groups (as determined by the Web site administrator).

- *Time window identification*: We retain the ordering in the page requests and we assume that a time difference of 30 min between two requests of the same user indicates different sessions, we end up with several sessions of the following form: $s_{ij} = 5\ 3\ 12\ 6\ 4\ 3\ 8\ 8\ 6\ 4\ 4\ 1$, where s_{ij} is the session j for user i, where its elements are the Web page categories.

Considering the above pre-processing of Web data, the session identification process can be formulated as follows (the variables notations are summarized in Table 2): Let $R^i = \{r_{i1}, \ldots, r_{iN_i}\}$ be the ordered list of i-th user access records in the log (sorted by the ascending order of the access frequency), and $t_{r_{ij}} (0 < j \leqslant N_i)$ be the time when r_{ij} was logged in the Web log file. Let $A = \{A_1, \ldots, A_V\}$ be the list of the categories and V the total number of categories. We assume that each r_{ij} is represented by one of these categories and the user i generates L_i sessions (possibly unequal length ordered sequences of pages). Assuming that W is the total number of users in the Web log file, let $D = \{R^1, \ldots, R^W\}$ be the observed access records for all the users in the log.

Definition 1 (*Web user session*). The sessions for an individual user i are defined as a list of subsets

$$s_{i1} = \{r_{i1}, \ldots, r_{in_1^i}\}, s_{i2} = \{r_{i(n_1^i+1)}, \ldots, r_{i(n_1^i+n_2^i)}\}, \ldots, s_{iL_i} = \{r_{i(n_1^i+\cdots+n_{L_i-1}^i+1)}, \ldots, r_{i(n_1^i+\cdots+n_{L_i}^i)}\}$$

where $t_{r_{i(n_1^i+1)}} - t_{r_{in_1^i}} \geqslant 30$ minutes, $N = n_1^i + \cdots + n_{L_i}^i$ and $s_{i1} \cup s_{i2} \cup \cdots \cup s_{iL_i} = S_i$. Then, the set $S = \{S_1, \ldots, S_W\}$ represents the sessions for all the users in our data set.

3.2. Determining the number of clusters

The proposed clustering approach assumes that the number of clusters is known at priori and as commonly employed in model-based schemes, the number of clusters might be determined by using several probabilistic values and methods, such as Bayesian Information Criterion (BIC), Bayesian approximations, or bootstrap methods (Fraley & Raftery, 1998). The present evaluation of the optimal number of clusters for our model is inspired from (Cadez et al., 2003) where a Bayesian approximation is used.

Formally, we assume that there are K clusters, denoted by $C_1, \ldots, C_K$, and each of them is generated from its own probability distribution. Once the model is specified, we use the EM algorithm and the probabilistic out-of-sample log likelihood evaluation to determine the best number of clusters. A model is fitted on a sub-

Table 2
Variables notation

Variable	Description
W	The total number of users
R^i	The ordered list of ith user access records in the Web log file
D	The set of all R^i
N_i	The total number of access records in R^i
r_{ij}	The jth access record (request) of user i
$t_{r_{ij}}$	The time when r_{ij} was logged in the Web log file
S	The sessions for all the users
S_i	The sessions for user i
L_i	The total number of sessions for user i
s_{ij}	The jth session for user i
n_i^j	The length for session s_{ij} (the number of access records that includes a session)
A_j	The jth state
V	The number of categories
C_k	The cluster k
θ^k	The set of the parameters for the cluster C_k
P	The transition matrix of a homogeneous ergodic Markov chain
K	The total number of clusters
$\text{Total\_sessions}_{C_k}$	The total number of sessions that belong to cluster C_k
f_j^k	The equilibrium probability of page category A_j in cluster C_k

sample of sessions (the so-called training data set) and then scored on the remaining data (the so-called testing data set). Thus, we get an objective measure of how well each model fits with the data. In order to determine the number of clusters, we choose the model with the minimum out-of-sample predictive log score for many values of K, i.e. we select that value for K, which minimizes the following equation:

$$\text{score}(K, Dtrain) = -\frac{\sum_{i=1}^{W}\sum_{j=1}^{L_i}\log_2 p(X = s_{ij}|\theta^K)}{\sum_{i=1}^{W}\sum_{j=1}^{L_i} n_j^i}, \tag{1}$$

where θ^K are the parameters obtained from EM algorithm (explained in the next subsection), and n_j^i is the length of session s_{ij}.

3.3. The clustering approach

Since the Web users' sessions have been identified, we cluster them by using a model-based clustering approach, the EM algorithm. Each resulted cluster contains a set of Web users' sessions generated by a probabilistic distribution. Each distribution is determined by a set of parameters which are different for each cluster. Specifically, the sessions in each cluster are represented by a first-order Markov ergodic chain. By the term "ergodic", we mean a Markov chain that has the following two properties: (1) Each node can reach any other node (all states intercommunicate), (2) the chain is not periodic (all states have period one). Such properties hold in the context of defining Web users' navigation behavior, since a Web user may navigate to every page, independently on which page had been previously visited and as proved in (Baldi et al., 2003) is periodic. In our framework, the parameters of each first-order Markov chain correspond to *an individual transition matrix* (which contains the transition probabilities among the states) and a *vector* (which represents the initial state probabilities). Thus, using a Markov chain model, we model the probability that the user will go to a certain page category given that he/she is viewing the current page category. Therefore, we have a transition matrix of size $V \times V$ (where V is the number of categories) and a set of V initial probabilities describing how likely is that user will begin a navigation session in a given page category. The following matrix $\mathbf{P}$ shows the structure of such a transition matrix where the probability for a user to navigate from page category A_i to page category A_j is denoted by P_{ij}.

$$\mathbf{P} = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_V \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \cdots\cdots & A_V \end{array} \\ \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1V} \\ P_{21} & P_{22} & \cdots & P_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ P_{V1} & P_{V2} & \cdots & P_{VV} \end{pmatrix} \end{array}$$

Definition 2 [Sessions Cluster]. The sessions are assigned to one of the underlying clusters by using the *hard assignment policy*. More specifically, a session s_{ij} belongs to cluster $C_k (1 \leqslant k \leqslant K)$ if and only if $p(x = s_{ij} \mid \theta^k) = \max\{p(x = s_{ij} \mid \theta^1), \ldots, p(x = s_{ij} \mid \theta^K)\}$, where θ^k is the set of the parameters for the cluster C_k.

From the above definition, it occurs that if the values of θ^k were observed, we would assign the users' sessions into clusters. However, these values are hidden. In order to learn the set of parameters θ^k for each cluster C_k, the Expectation-Maximization (EM) algorithm is used. The EM algorithm originates from (Dempster, Laird, & Rubin, 1977) while in (Cadez et al., 2003) a method for employing EM on users' sessions is proposed. The EM algorithm searches for a maximum likelihood hypothesis by repeatedly re-estimating the expected values of the hidden variables θ^k given a current hypothesis. Specifically, the following steps are repeated:

- *The expectation E-step*: Given a set of parameter estimates, the E-step calculates the conditional expectation of the complete-data log likelihood given the observed data and the parameter estimates.

- *The maximization M-step*: Given a complete-data log likelihood, the M-step finds the parameter estimates to maximize the complete-data log likelihood from the E-step.

The two steps are iterated until convergence, based on the core idea of the EM approach that the current hypothesis is used to estimate the unobserved variables, and the expected values of these variables are then used to calculate an improved hypothesis.

In terms of the complexity of the EM algorithm, it depends on the complexity of the E and M steps at each iteration (Dempster et al., 1977). For example, in our case (Markov mixtures) the complexity is linear in the sum of the lengths of all sessions, whereas in more complex mixture models the complexity can be higher.

4. Validating web users' sessions clusters

In this section we present a novel validation method, which evaluates the model-based clustering schemes. The approach belongs to the internal type since the clustering result is evaluated in terms of quantities obtained from the data set itself.

The first stage of the procedure takes as input the resulted clusters, where each one consists of users' sessions. As we have already mentioned, the objects of each cluster are assumed to follow a first-order Markov ergodic model[4]. The main idea is to consider the *equilibrium distribution* of the transition matrices for each cluster. These distributions represent the probabilities of a user to access each state in infinite number of states, independently of its initial state. The reason we use the equilibrium distribution is that it offers an *effective* and *objective* view for the navigation behavior of Web users, since it provides a strong "*long-term*" indication for the most popular Web pages.

Theorem 1. *If* $\mathbf{P}$ *is the* $V \times V$ *transition matrix of a homogeneous ergodic Markov chain, then there is a unique vector* $\mathbf{f} = (f_1, \ldots, f_V)$, *such that*

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \mathbf{f} \\ \mathbf{f} \\ \vdots \\ \mathbf{f} \end{pmatrix} = \begin{pmatrix} f_1 & \ldots & f_V \\ f_1 & \ldots & f_V \\ \vdots & \ddots & \vdots \\ f_1 & \ldots & f_V \end{pmatrix} \tag{2}$$

Proof. A thorough study and classification of finite Markov chains and the proof of this theorem is given in (Cox & Miller, 1997).

This theorem offers us a way of approximately evaluating the access frequencies of the nodes (visited categories in our case), by simply calculating powers of the transition matrix. It gives us a way to evaluate the relative*frequency* of accessing (retrieving) nodes $1, \ldots, V$ respectively in a long run, based on the transition probabilities of the initial browsing graph. It is known that in the theory of stochastic processes the vector f is called the equilibrium or stationary distribution of the Markov chain since any element represents the limiting probability of accessing the respective nodes $1, \ldots, V$ after infinite number of steps.

The next step involves the representation of each cluster by its corresponding equilibrium distribution and the clustering validation, performed by testing the homogeneity of the cluster equilibrium distributions using the χ^2 test.

In general the χ^2 test is used for testing the independence of two categorical variables or alternatively the distribution homogeneity in the categories of one variable with respect to the other (Snedecor & Cochran, 1989). The χ^2 is a statistic, i.e. a quantity computed from observations, which is used to measure the dissimilarity among probability distributions. Considering therefore that each cluster can be represented by a

[4] In view of this, it should be emphasized that the following method is valid only if all clusters can be represented by ergodic Markov chains. However, in a large number of experiments conducted by the EM algorithm, it has been observed that the above condition is always satisfied (Baldi et al., 2003; Cadez et al., 2003).

Table 3
A contingency table for clusters and page categories

	Page categories				
Clusters	A_1	A_2	...	A_V	Row sums
C_1	O_{11}	O_{12}	...	O_{1V}	Y_1
C_2	O_{21}	O_{22}	...	O_{2V}	Y_2
...	...	...	...	...	...
C_K	O_{K1}	O_{K2}	...	O_{KV}	Y_K
Column sums	X_1	X_2	...	X_V	Sum

probabilistic distribution (the equilibrium distribution), we can directly apply this homogeneity test. So in our case, we essentially consider two variables: one having as values the K derived clusters and the other having as values the V page categories. These variables are cross-tabulated in a *contingency table*, i.e. a table summarizing the relation between clusters and page categories. The general form of such a cross-tabulation is given in Table 3. The contingency table is next used for the computation of the χ^2 statistic and the subsequent test of homogeneity.

In order to describe in detail the procedure, we assume that the clustering algorithm results in K clusters denoted by $C_1,\ldots,C_K$. The V different page categories, denoted by $A_1,\ldots,A_V$, are distributed in cluster k according to its equilibrium probability distribution denoted by $\mathbf{f}^k = (f_1^k,\ldots,f_V^k)$. Each *observed frequency* O_{ij} in the cells of the contingency table (Table 3) is computed by multiplying the equilibrium probability f_j^i of the A_j page category with the number of sessions that belong to cluster C_i (denoted by $\text{total\_sessions}_{C_i}$).

Having formed the contingency table, our aim now is to test the homogeneity of the clusters with respect to the distribution of the page categories in each of them. If the test shows significant heterogeneity, this can be attributed to the ability of the clustering algorithm to produce distinguishable groups. Moreover, this heterogeneity can be further analyzed in order to reveal and interpret the characteristics of the users' behavior in each cluster. On the other hand, if the test shows that the clusters are homogeneous (i.e. the categories are distributed more or less similarly) we can infer that the clustering was not successful, and the clusters cannot be interpreted.

The χ^2 statistic is computed from the contingency table by the following formula:

$$\chi^2 = \sum_{i=1}^{K}\sum_{j=1}^{V}\frac{\left(O_{ij} - Y_i \times \frac{X_j}{\text{Sum}}\right)^2}{Y_i \times \frac{X_j}{\text{Sum}}}, \tag{3}$$

where

$$O_{ij} = f_j^i \cdot \text{total\_sessions}_{C_i}, \quad X_j = \sum_{j=1}^{V} O_{ij}, \quad Y_i = \sum_{i=1}^{K} O_{ij}, \quad \text{and} \quad \text{Sum} = \sum_{i=1}^{K}\sum_{j=1}^{V} O_{ij}$$

and it is used to test the null hypothesis that the distributions of the page categories in each cluster are not significantly different. A large value of the χ^2 criterion shows that the equilibrium distributions of the clusters are significantly different, which in turn is an indication of the heterogeneity among clusters. In order to judge whether χ^2 is really large, we need to know a critical value for the boundary of the area of hypothesis's rejection. In order to find this critical value, we should define the level of significance a (probability of erroneously rejecting the null hypothesis) and the degrees of freedom (df). From statistical theory we know that under the null hypothesis of homogeneity, the χ^2 statistic has asymptotically a χ^2 distribution with $(K-1)\times(V-1)$ degrees of freedom. Thus, if the value of χ^2 statistic, computed from our data, is greater than a critical value of the χ^2 distribution corresponding to probability a, denoted by $\chi^2_{(K-1)\times(V-1);a}$, we reject the null hypothesis at the level of significance a. In such a case we can conclude that the clustering algorithm produced separable groups.

5. Interpreting users' sessions clusters

From the validation procedure of the previous section we obtain a strong indication about the overall dissimilarity of the clusters, using the χ^2 statistic. However, some clusters may be closer than others while some

categories may be highly associated with certain clusters. These relations cannot be investigated by the χ^2 test and therefore a further analysis of clusters is essential in order to reveal and interpret certain associations.

Interpreting the navigation behaviors exhibited by the Web users' sessions in each cluster is important for a number of tasks, such as providing of valuable insight about users' preferences, designing of a Web site, identifying malicious visitors and managing targeted advertising. It also helps in understanding the sessions of different users' groups and, therefore, in organizing the Web site to better suit the users' needs. Furthermore, interpreting the results of clusters contributes in identifying and providing customized services and recommendations to Web users by exploring relations between the categories. However, the interpretation of clusters is a difficult and time-consuming process due to large-scale data sets and its complexity. To address this interpretation problem the research community has focused on visualization approaches (Baldi et al., 2003). Clustering visualization can help the Web administrators to visually perceive the clustered results, and sometimes oncover hidden patterns in data.

In this section, we introduce a novel clustering interpretation approach by analyzing the contingency table, which has been constructed for the validation process described in the previous section. The analysis uses the statistical methodology known as correspondence analysis method (CO-AN).

The main goal of CO-AN is to describe the relationships between two categorical variables in a contingency table. These relationships are described by projecting the values of the variables as points on a two-dimensional space, in such a way that the resulting plot describes simultaneously the relationships between the categories of each variable. For each variable, the distances between points in the plot reflect the relationships between the categories. Similar categories are plotted close to each other while distant points show dissimilarity. The computation of the coordinates in the two-dimensional axis system are based on the χ^2 statistic as measure of distance. Mathematical details of CO-AN can be found in (Johnson & Wichern, 1998).

In our case we can apply the CO-AN method to the rows and columns of Table 3 in order to explore further the relationships between clusters and page categories. The obtained graphical representation provides a meaningful interpretation of clusters and therefore useful information regarding users' navigation behaviors. Consider for example the case where a Web developer wants to arrange the structure of a site such as to interlink associated pages. This can be achieved by the proposed method which finds such associations.

At this point we have to emphasize that the validation and the interpretation procedures do not concern only a specific clustering result, i.e. the one obtained from the determination of the optimal number of clusters according to the BIC criterion and the subsequent application of the EM algorithm. If we determine another number of clusters (empirically or by setting another criterion) or if we apply a different algorithm, the resulting groups will be probably quite different. However, the suggested procedure should be applied to any grouping so as to examine whether the groups are separable and interpretable.

6. Experimentation-results

6.1. The data sets

The methods described can be applied on usage data of any Web site. In this paper, they have been applied on two real data sets: the first data set (called msnbc data set) comes from an active popular commercial Web server (msn.com[5]) which consists of a daily record of approximately 6000 users' sessions, with an average of 5.7 page views per session. This data set includes visits which are recorded at the level of URL category and are recorded in time order and no pre-processing was required since data set was given in sessions. The second data set (called csd data set) comes from an educational Web server (the Department of Computer Science in Aristotle University of Thessaloniki), which consists of approximately 3437 users' sessions, with an average of 3.3 page views per session. Sessions in this data set originate from a log file of 1,000,000 records and it refers to the categories assigned to the total of 11,342 Web pages. Table 4 summarizes the details of these data sets.

Each event in the sequence-session corresponds to a user's request for a page, which is recorded at the level of page category and not at the level of URL. In order to apply the methods discussed in the previous sections

[5] Msnbc.com anonymous Web data: http://kdd.ics.uci.edu/databases/msnbc/msnbc.html.

Table 4
Data sets details

Data set	Time period	Number of sessions	Web pages per session
Csd	1/4/2003–1/11/2003	3437	3.3
Msnbc	9/28/1999	6000	5.7

we have to use transition matrices to represent the navigation steps of the users from category to category. Since it is important for the interpretation of the clustering to know the first and the last page categories, we defined two auxiliary categories: the "start-state" and the "end-state". Although it is sufficient for the construction of the transition matrices to define only one auxiliary category (for example the "outer-state") we choosed to work with two, such that in the interpretation phase we can investigate which page categories are associated with "entry" and "exit" by plotting them separately with distinct points.

The categories of each data set are described in Tables 5 and 6. Then, for each data set, we select the first (ordering by time) of the 70% of the total sessions as training data set and the rest as testing data set in order to determine the number of clusters.

6.2. Clustering and validating of users' sessions

After preprocessing and setup of the datasets, we determine the number of clusters by finding the value that minimizes the out-of-sample predictive score, given in Eq. (1). Figs. 3 and 4 show several out-of-sample log-likelihoods for varying number of clusters. The *x-axis* represents the number of clusters, while the *y-axis* represents the out-of-sample log-likelihood. From these figures, it is evident that the out-of-sample log-likelihood is minimized when the number of clusters is 5 and 4 for the *msnbc* and the *csd* data set, respectively.

The next step is to cluster the users' navigation sessions as described in Section 4. Each cluster is represented by a probabilistic distribution (first-order Markov model). Using the EM algorithm we learn the parameters of each Markov model as well as the proportion of users' sessions assigned to each cluster. Then, we assign each user session to a cluster according to Definition 2 and once the clusters have been identified, the next step is to validate them by forming the corresponding contingency table and using the χ^2 test on their equilibrium distribution. mfloatTable 7Tables 8 and 9 present the contingency tables derived after the clustering of the two data sets. Since the frequencies in the cells are computed by Eq. (3) we give the tables with rounded entries. Table 7 presents the results of χ^2 test at the level of significance $a = 0.001$ for the *msnbc* and the *csd* data sets. As depicted in this table, for both data sets, the value of the χ^2 statistic is much higher than the critical value $\chi^2_{(K-1)\times(V-1);a}$. From this result we can conclude that the clustering algorithm gave well separated and distinguishable clusters.

Table 5
Web page categories (msnbc data set)

A_1	Start-state	A_6	Opinion	A_{11}	Health	A_{16}	Summary
A_2	Frontpage	A_7	On-air	A_{12}	Living	A_{17}	Bbs
A_3	News	A_8	Misc	A_{13}	Business	A_{18}	Travel
A_4	Tech	A_9	Weather	A_{14}	Msn-sports	A_{19}	End-state
A_5	Local	A_{10}	Msn-news	A_{15}	Sports		

Table 6
Web page categories (csd data set)

A_1	Start-state	A_5	Personnel	A_9	Links
A_2	Home	A_6	Labs	A_{10}	Misc
A_3	Information	A_7	Students	A_{11}	End-state
A_4	Studies	A_8	Conferences		

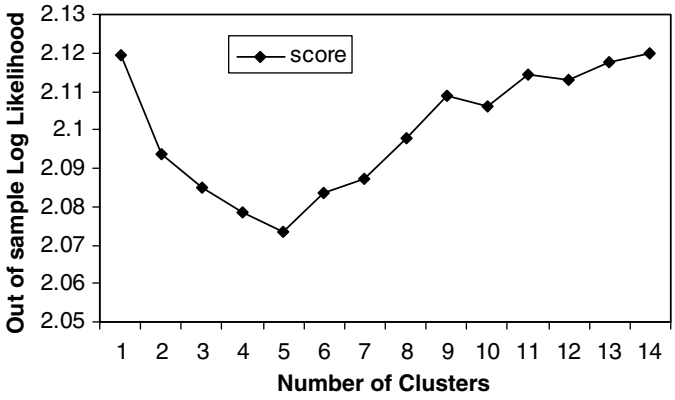


Fig. 3. Number of clusters (msnbc data set).

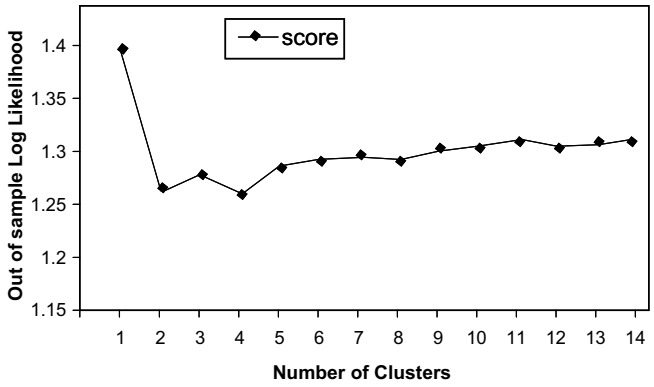


Fig. 4. Number of clusters (csd data set).

Table 7
Chi-square testing

	χ^2	$\chi^2_{(K-1)\times(V-1);a}$
Msnbc data set ($K=5$, $V=19$, $a=0.001$)	1795	115
Csd data set ($K=4$, $V=11$, $a=0.001$)	578	60

Table 8
A contingency table test (msnbc data set)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	*Sum*
C_1	12	42	91	29	24	18	95	120	36	29	58	39	31	24	86	17	13	15	107	886
C_2	33	38	47	49	72	36	73	37	37	42	38	38	131	38	41	42	34	34	188	1049
C_3	19	328	140	65	79	169	45	77	28	30	51	44	75	19	115	25	33	27	180	1548
C_4	22	174	57	56	36	29	74	76	40	70	28	28	25	22	25	36	22	23	158	1000
C_5	26	50	48	35	235	32	39	62	225	80	29	41	46	145	83	47	26	31	238	1517
Sum	112	632	383	234	445	283	326	371	366	250	205	189	308	249	350	167	128	130	871	6000

6.3. Cluster analysis interpretation

After the validation procedure, the next step is to apply the CO-AN method to the contingency tables in order to visualize the clusters and the page categories.

Table 9
A contingency table test (csd data set)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	*Sum*
C_1	5	94	10	43	36	91	8	6	5	5	27	330
C_2	12	173	16	16	16	71	15	31	12	13	53	428
C_3	74	95	78	91	82	221	79	91	74	76	363	1324
C_4	70	178	78	72	102	94	74	126	70	71	420	1355
Sum	161	540	182	222	236	477	176	254	161	165	863	3437

To visualize the quality the of clustering algorithm, we depict the associations among the clusters for both data sets. As can be seen in Figs. 5 and 6, each cluster is represented by a point. According to these plots, a general observation that can be taken for both data sets is that the resulted clusters are separable since there are no coincident points. It is worth mentioning that this observation is in accordance with the results of the χ^2 test that we have been obtained previously. An inside view of each cluster will be presented in Section 6.3.2.

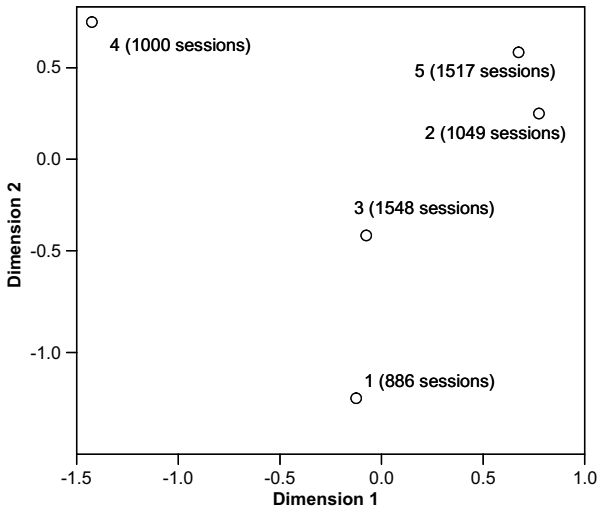


Fig. 5. Correspondence map of clusters (msnbc data set).

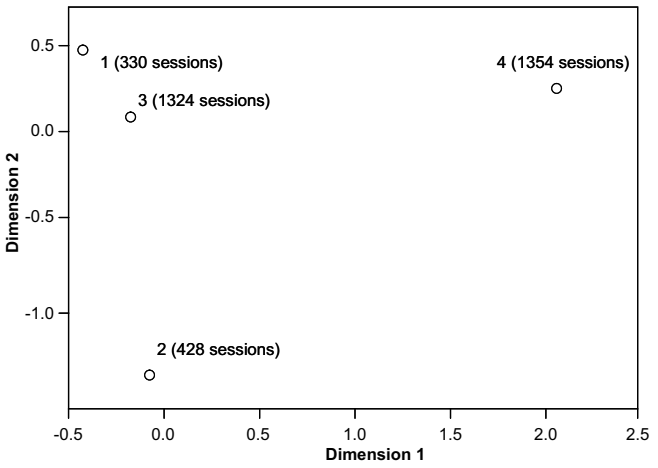
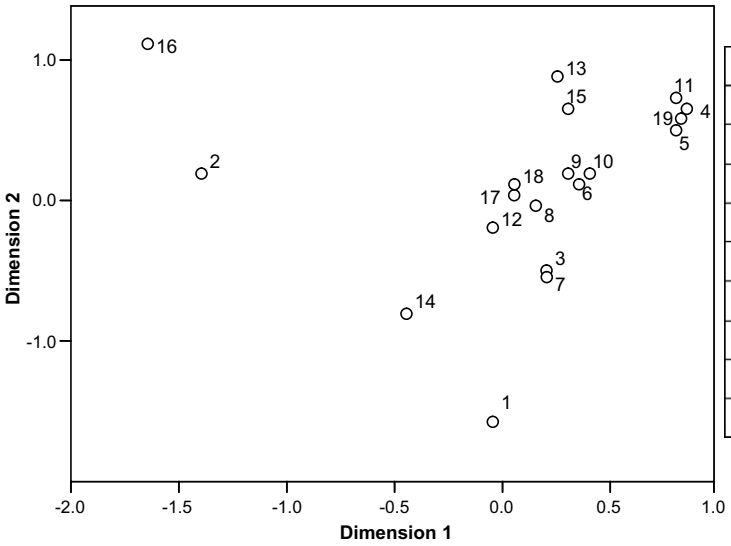


Fig. 6. Correspondence map of clusters (csd data set).

6.3.1. Visualizing the associations of web pages

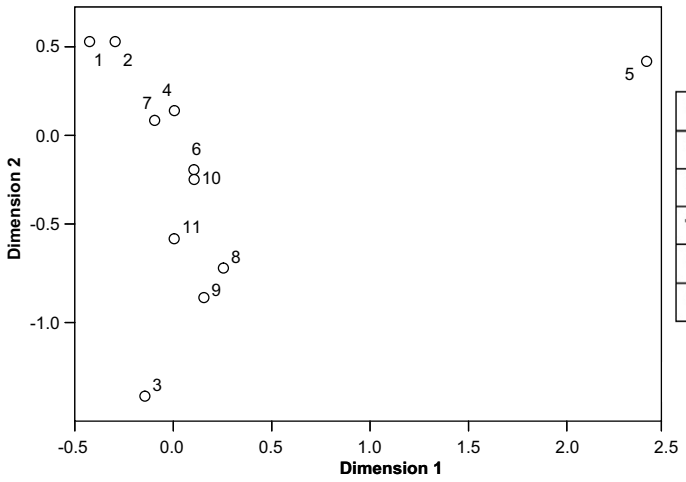
The proposed clustering analysis can also be used to graphically display the usage relevance among Web page categories, since Web requests are recorded at the level of page category (in practice, categories are typically determined by the Web site administrator). Figs. 7 and 8 illustrate the usage associations between Web page categories as resulted from the employed correspondence analysis for the *msnbc* and *csd* data sets, respectively. Each category is represented by a point, i.e. if some points are close to each other, this means that the Web page categories (corresponding to these points), are associated with each other. Thus, observing these figures, we can extract useful information and understand Web users' navigational behavior and trends. The following observations in Figs. 7 and 8 results will help in understanding the web usage trends:

- *Msnbc data set*: From Fig. 7, it is evident that the Web users who visit pages about "News", most probably will visit, in the same navigation session, pages about "On air" (since categories 3 and 7 are closely associated). Similarly, pages about "Bulletin board service (bbs)" are closely associated with pages about



1	Start-state	11	Health
2	Frontpage	12	Living
3	News	13	Business
4	Tech	14	Msn-sports
5	Local	15	Sports
6	Opinion	16	Summary
7	On-air	17	Bbs
8	Misc	18	Travel
9	Weather	19	End-state
10	Msn-news		

Fig. 7. Correspondence map of Web page categories (msnbc data set).



1	Start-state	7	Students
2	Home	8	Conferences
3	Information	9	Links
4	Studies	10	Misc
5	Personnel	11	End-state
6	Labs		

Fig. 8. Correspondence map of Web page categories (csd data set).

"travel" (since categories 17 and 18 are closely associated). Furthermore, the users tend to exit from a Web navigation path, when they have previously visited Web pages about "Tech", "Local" or "Health", i.e. there is a trend to follow a particular navigation pathway prior exiting. This fact is probably an indication that these pages are not quite attractive and the users abandon the Web site. Another remark is that when users visit pages either about "Frontpage" or about "Summary", they do not visit pages about "Health" (since categories 2 and 16 are quite loosely associated). Moreover, it should be noted that the "Start-state" (category 1) is too far from the "End-state" (category 19), which means that these states are not associated with each other. This result justifies are initial choice to consider a different state for the beginning and the ending of each session.

- *Csd data set*: From Fig. 8, it is evident that the "start state" and "Home" are closely associated with each other, i.e. there is an indication that the first page that Web users visit in this Web site is the "Home" page. Also it seems that users tend to visit in the same sessions categories in couples such as categories "Conferences" and "Links" and categories "Labs" and "Misc". On the other hand, it is evident that the category "Personnel" is visited solely and separately than other categories. This is due to the potential tend to search for particular people individually, since it is quite often to navigate on the Web when looking for particular academics or faculty (e.g. in an effort to look for a collaboration or establishing a communication contact with a professor).

6.3.2. An inside view of the clusters

Except of analyzing the association among Web page categories it is also very useful to have a view about the contents of each cluster, since a deeper knowledge for the inside of each cluster can draw useful and meaningful inferences for the users' navigation behavior. More specifically, Figs. 9 and 10 depict the percentage frequency of requested Web page categories observed in each cluster to help in understanding users' navigation behavior for both *msnbc* and *csd* data sets. The following comments aim at giving an inside view of the resulted clusters per data set:

- *Msnbc data set*: From Fig. 9, it is evident that the users' sessions that belong to cluster 1 refer to of a wide range of pages, showing more preference to "Sports", "Misc", "Local" and "News" categories. This may be an indication that the users who navigated as described in these sessions, had no particular interest at a specific category but they have rather showed browsing behavior within the Web site, navigating over various page categories. On the other hand, the users' sessions in cluster 2 show a special interest to "Business", and "On-air", i.e. these users are probably business-oriented (e.g. they are interested in stock

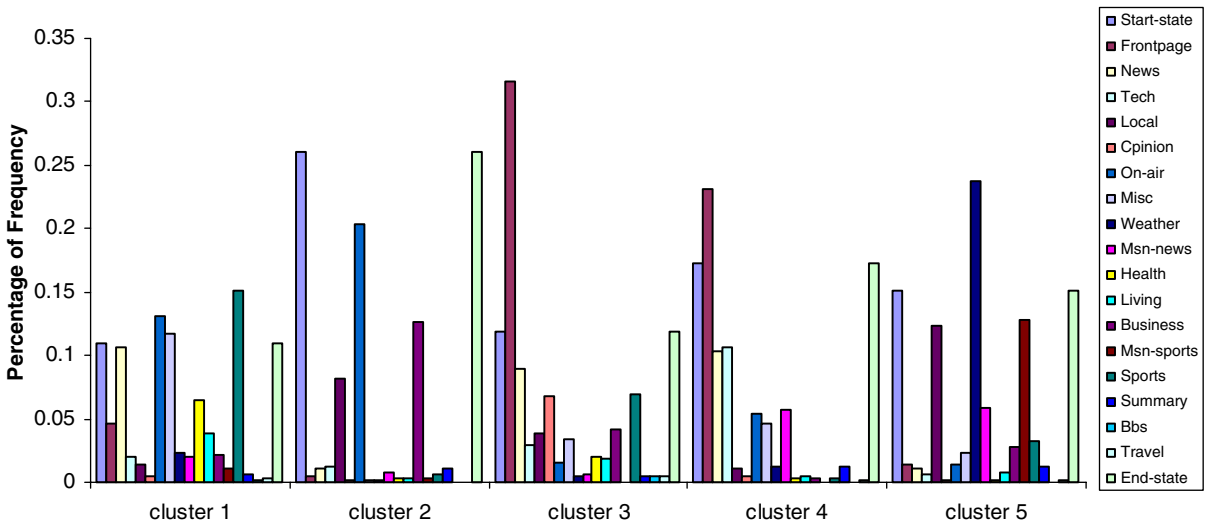


Fig. 9. The percentage frequency of Web page categories for each cluster (msnbc data set).

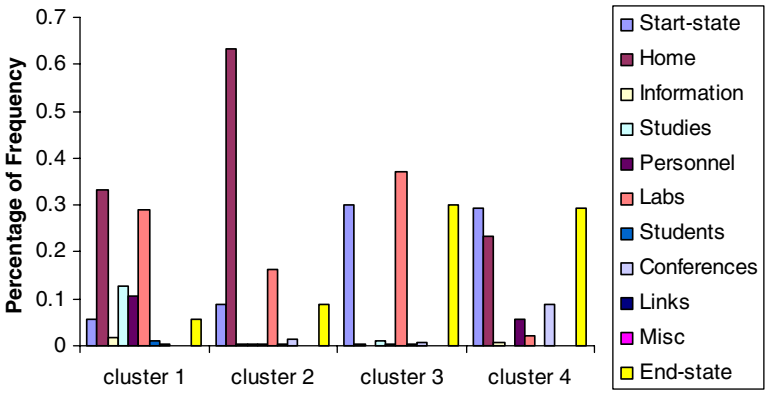


Fig. 10. The percentage frequency of Web page categories for each cluster (csd data set).

market which involved both business and it might be shown on communications media). In cluster 3, 30% of the total requested Web page categories belong to "Frontpage", i.e. we might have users interested in Web page authoring. Users' sessions on cluster 4 show high interest for the "Frontpage", "News" and "Tech" categories (and low interest for all the other categories), i.e. it might declare focused navigation on using technology for Web news site authoring. On the other hand, users' sessions in cluster 5 present high interest only for the categories which refer about "Weather", "Local" and "Msn-sports" so they might refer to users interested in a game and they want to get information about local weather predictions.

- *Csd data set:* The users' sessions in cluster 1 show a special interest to "Home", and "Labs", i.e. users who followed such sessions might probably be (under-) graduate students who check the Home page and the Lab schedule to check whether there are any announcements about their lab courses. In clusters 2 and 4, the most popular page category is also the "Home", i.e. we might conclude that we have users who are interested in the csd department establishment. Finally, in cluster 3, the users' sessions show high interest for the "Labs" and low interest for all the other categories, i.e. we have an indication that these users are mostly (under-) graduate students who have focused interests on their Lab assignments, with no further interest for any other information about the department. Based on these observations an overall conclusion for the particular dataset is that users focus on navigating on few categories, namely the "Home" and "Lab" categories.

6.3.3. Exploiting cluster validation and interpretation in practice

Visualizing the associations and then interpreting clusters of Web user sessions under the proposed procedures offers important information as discussed in the earlier subsections. Evaluating both the proposed validation and interpretation approaches on our considered real data datasets originating from two Web servers, we noticed that there are plenty of inferences that may be drawn in terms of users' preferences, interests and origin. In this context, the adoption of the proposed validation and interpretation procedures in current Web-related applications might offer important benefits. Some indicative such applications are the following:

- *E-commerce applications*: the proposed procedures in conjunction with a corresponding e-commerce data analysis (Kohavi, Mason, Parekh, & Zheng, 2004), may offer important knowledge about customers origin, needs and preferences. For example, the use of the interpretation procedure can maximize the sales by minimizing the route of the potential customers' page visits from homepage to the requested (for purchase) product. Moreover, validation of clusters of users' sessions may indicate the separation in customer habits or product (dis)likes so that the underlying commercial company might guide certain advertising tasks towards a particular cluster, i.e. a group of customers with common navigational behavior.
- *Web site administration*: the proposed approaches are also beneficial in terms of realizing how and with which patterns the Web site page categories are visited, so to take certain actions for revising Web site's structure and presentation, in relation to Web site evaluation or reorganization (Chakrabarti, 2003). There-

fore, based on the results of visualizing certain associations among page categories, a Web site administrator might decide to take certain site structure rearrangements so that the accessing speed and the user interaction with the site will be improved.
- *Caching and prefetching*: having the proposed validated clusters of sessions which have been proven to show clearly separated groups, actions of caching and/or prefetching clusters may be beneficial in terms of the users' perceived latency. The proposed validation and interpretation procedures are in accordance with the need to deliver the appropriate content to the interested users in a timely, scalable, and cost-effective manner (Pallis & Vakali, 2006).

Certainly, there are more applications may benefited from the proposed work, such as searching on the Web, as well as personalization and recommendation engines, which are also demanding in terms of Web usage interpretation and understanding.

7. Conclusions

This paper presents a complete framework for model-based cluster analysis for Web users' sessions. Taking into consideration that the Markov models may provide valuable information for users' navigation behavior which is often hidden, it is crucial to discover the hidden meaningful relationships among users' sessions as well as between users' sessions and Web objects. Towards this direction, the proposed validation and interpretation methods have been proved efficient and very robust for validating Web users' sessions clusters as well as inferring meaningful results from these clusters. Specifically, the validation procedure is a newly presented approach in the literature for validating model-based clusters and the interpretation procedure is a novel visualization method for interpreting the clustering results by revealing interesting features for Web users' navigation behavior and their interaction with the content/structure of Web sites.

Evaluating both the proposed validation and interpretation approaches on real data originating from intensive Web servers, we noticed that the proposed approach is a valuable tool for various Web-based applications. Such indicative applications are benefited since Web usage clusters validation and interpretation is:

- facing some of the Web administrators problems, who need to validate and interpret the resulted clusters in order to improve site's structure and organization;
- dealing with customers characterization in e-commerce applications, so that a company might increase their promotion or marketing actions at specified customer groups;
- identifying appropriate clusters which may then be cached or prefetched at particular locations. Moreover these clusters might guide certain searching and recommendation tasks.

It is interesting to investigate in a future work the impact of the proposed validation and interpretation on particular applications. In this context, it is challenging to consider both validation and interpretation at current application testbeds in order to assess which of the two procedures (validation or interpretation) is most effective and beneficial.

Acknowledgements

The authors appreciate and thank the anonymous reviewers for their valuable comments and suggestions, which have considerably contributed in improving the paper's content, organization and readability.

References

Anderson, C. R., Domingos, P., & Weld, D. (2002). Relational Markov models and their application to adaptive web navigation. In *Proceedings of the 8th international conference on knowledge discovery and data mining* (pp. 143–152). New York: ACM.

Baldi, P., Frasconi, P., & Smyth, P. (2003). *Modeling the Internet and the Web*. New York: Wiley.

Banerjee, A., & Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. In *Proceedings of the workshop on web mining, SIAM conference on data mining* (pp. 33–40). Chicago, USA.

Berendt, B., & Spiliopoulou, M. (2000). Analysis of navigation behaviour in Web sites integrating multiple information systems. *VLDB Journal, 9*(1), 56–75.

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

Catledge, L., & Pitkow, J. (1995). Characterizing browsing behaviors on the world wide web. *Computer Networks and ISDN Systems, 6*(27), 1065–1073.

Chakrabarti, S. (2003). *Mining the Web*. San Fransisco: Morgan Kaufman.

Chen, Z., Fu, A., & Tong, F. (2003). Optimal algorithms for finding user access sessions from very large web logs. *World Wide Web: Internet and Information Systems, 6*, 259–279.

Chen, K., & Liu, L. (2003). Validating and Refining Clusters via Visual Rendering. In *Proceedings of the 3rd International Conference on Data Mining (ICDM 2003)* (pp. 501–504). Melbourne, Florida: IEEE.

Chen, M. S., Park, J. S., & Yu, P. S. (1998). Efficient data mining for path traversal patterns. *IEEE Transactions on Knowledge and Data Engineering, 10*(2), 209–221.

Cox, D. R., & Miller, H. D. (1997). *The theory of stohastic processes*. New York: Chapman and Hall.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B, 39*, 1–22.

Desphpande, M., & Karypis, G. (2001). Selective Markov models for predicting web page accesses. In *Proceedings of the 1st SIAM conference on data mining*. San Diego, USA.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for Web personalization. *ACM Transactions on Internet Technology (TOIT), 3*(1), 1–27.

Fraley, C., & Raftery, A. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal, 41*(8), 578–588.

Fu, Y., Sandhu, K., & Shih, M. Y. (1999). Clustering of Web users based on access patterns. In *Proceedings of the international workshop on Web usage analysis and user profiling, (WEBKDD 1999)*. San Diego, California: Springer-Verlag.

Goker, A., & He, D. (2000). Analysing Web search logs to determine session boundaries for user-oriented learning. In *Proceedings of the international conference of adaptive hypermedia and adaptive Web-based systems (AH2000)* (pp. 319–322). Trento, Italy: Springer-Verlag.

Gomory, S., Hoch, R., Lee, J., Podlaseck, M., & Schonberg, E. (1999). Analysis and visualization of metrics for online merchandizing. In *Proceedings of the 1st international workshop on Web usage analysis and user profiling (WEBKDD 1999)*. San Diego, USA: Springer-Verlag.

Govaert, G. & Nadif, M. (in press). Clustering of contingency table and mixture model. *European Journal of Operational Research*.

Gunter, S., & Bunke, H. (2003). Validation indices for graph clustering. *Pattern Recognition Letters, 24*(8), 1107–1113.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*(2–3), 107–145.

Hay, B., Vanhoof, K., & Wets, G. (2001). Clustering navigation patterns on a website using a sequence alignment method. In *Proceedings of 17th International Joint Conference on Artificial Intelligence*. Seattle, Washington, USA, 2001.

He, D., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic Web session identification. *Information Processing and Management, 38*(5), 727–742.

Huang, Z., Ng, M. K., & Cheung, D. W.-L. (2001a). An empirical study on the visual cluster validation method with fastmap. In *Proceedings of the 7th international conference on database systems for advanced applications (DASFAA 2001)* (pp. 84–91). Hong Kong, China: Springer-Verlag.

Huang, Z., Ng, M. K., Cheung, D. W.-L., & Ching, W. (2001b). A cube model for web access sessions and cluster analysis. In *Proceedings of the 3rd international workshop on mining web data (WEBKDD 2001)* (pp. 48–67). San Francisco: Springer-Verlag.

Huang, X., Peng, F., An, A., & Schuurmans, D. (2004). Dynamic web log session identification with statistical language models. *JASIST, 55*(14), 1290–1303.

Ibrahimov, O., Sethi, I., & Dimitrova, N. (2002). The performance analysis of a Chi-square similarity measure for topic related clustering of noisy transcripts. *Proceedings of the 16th international conference on pattern recognition* (Vol. 4, pp. 285–288). Quebec, Canada: IEEE Press.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys, 31*(3), 264–323.

Javed, Y., & Bhatti, A. I. (2005). Emitter recognition based on modified X-means clustering. In *Proceedings of the IEEE symposium on emerging technologies* (pp. 352–358). Islamabad: IEEE Press.

Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate statistical analysis*. Upper Saddle River: Prentice-Hall.

Kohavi, R., Mason, L., Parekh, R., & Zheng, Z. (2004). Lessons and challenges from mining retail e-commerce data. *Machine Learning, 57*(1–2), 83–113.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Journal of Marketing Science, 23*(4), 579–595.

Morey, L. C., & Agresti, A. (1984). The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement, 44*, 33–37.

Pallis, G., Angelis, L., & Vakali, A. (2005). Model-based cluster analysis for Web users sessions. In *Proceedings of the 15th international symposium on methodologies for intelligent systems (ISMIS 2005)* (pp. 219–227). Saratoga (NY) USA: Springer-Verlag.

Pallis, G., Angelis, L., Vakali, A., & Pokorny, J. (2004). A probabilistic validation algorithm for Web users' clusters. In *Proceedings of the IEEE international conference on systems, man and cybernetics (SMC 2004)*. Hague, Holland: IEEE.

Pallis, G., & Vakali, A. (2006). Insight and perspectives for content delivery networks. *Communications of the ACM, 49*(1), 101–106.

Roussinov, D. G., & Chen, H. (2001). Information navigation on the web by clustering and summarizing query results. *Information Processing and Management, 37*(6), 789–816.

Sen, R., & Hansen, M. H. (2003). Predicting a Web user's next request based on log data. *Journal of Computations Graph Statistics, 12*(1).

Shahabi, C., Faisal, A., Kashani, F. B., & Faruque, J. (2000). INSITE: a tool for interpreting users? Interaction with a Web space. In *Proceedings of the 26th international conference on very large data bases (VLDB 2000)* (pp. 635–638). Cairo, Egypt.

Shahabi, C., Zarkesh, A. M., Adibi, J., & Shah, V. (1997). Knowledge discovery from users Web page navigation. In *Proceedings of the 7th international workshop on research issues in data engineering (RIDE)*. Birmingham, England: IEEE.

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.

Snedecor, G., & Cochran, W. (1989). *Statistical methods* (8th ed.). Iowa State University Press.

Wang, W., & Zaiane, O. R. (2002). Clustering Web sessions by sequence alignment. In *Proceedings of the 13th international workshop on database and expert systems applications (DEXA 2002)*. Aix-en-Provence, France: Springer-Verlag.

Ypma, A., & Heskes, T. (2002). Categorization of Web pages and user clustering with mixtures of hidden Markov models. In *Proceedings of the 4th international workshop on mining Web data (WEBKDD 2002)* (pp. 31–43). Edmonton Alberta, Canada: Springer-Verlag.