Int. J. Knowledge and Web Intelligence, Vol.

A fuzzy bi-clustering approach to correlate web users and pages

Vassiliki A. Koutsonikola* and Athena Vakali

Department of Informatics,
Aristotle University,
Thessaloniki 54124, Greece
E-mail: vkoutson@csd.auth.gr
E-mail: avakali@csd.auth.gr
*Corresponding author

Abstract: With the rapid development of information technology, the significance of clustering in the process of delivering information to users is becoming more eminent. Especially in the web information space, clustering analysis can prove particularly beneficial for a variety of applications such as web personalisation and profiling, caching and prefetching and content delivery networks. In this paper, we propose a bi-clustering approach, which identifies groups of related web users and pages. The proposed approach is a three-step process that relies on the principles of spectral clustering analysis and provides a fuzzy relation scheme for the revealed users' and pages' clusters. Experiments have been conducted on both synthetic and real datasets to prove the proposed method's efficiency and reveal hidden knowledge.

Keywords: web users; web pages; fuzzy bi-clustering; spectral analysis.

Reference to this paper should be made as follows: Koutsonikola, V.A. and Vakali, A. (2009) 'A fuzzy bi-clustering approach to correlate web users and pages', *Int. J. Knowledge and Web Intelligence*, Vol.

Biographical notes: Vassiliki A. Koutsonikola received the BS in Computer Science from Aristotle University of Thessaloniki, Greece in 2001, the MS in Information Systems from the University of Macedonia, Greece in 2003 and the PhD from Aristotle University of Thessaloniki in January 2009. Her research interests include clustering, directory services and network-based data organisation.

Athena Vakali received a BSc in Mathematics, a PhD in Informatics from Aristotle University of Thessaloniki (AUTH), Greece and a MSc in Computer Science from Purdue University, USA. Currently, she is an Associate Professor at the Department of Informatics, AUTH, and the head of the operating systems web/internet data sources management research group. Her research interests include various aspects and topics of the web information systems, such as web data management, content delivery on the web, web data clustering, web caching, web 2.0 data mining, XML-based authorisation models, text mining and multimedia data management.

1 Introduction

The World Wide Web has rapidly emerged as a popular medium, which enables massive information publishing and retrieval. However, its huge growth has led to an information overload that continuously expands causing various problems to web users. These problems are usually related to the accuracy of the retrieved information, which is characterised by low precision or irrelevance (Liu, 2007). Moreover, the delivered web information lacks personalisation while the requested web pages do not conform with users preferences (Eirinaki and Vazirgiannis, 2003).

Web clustering is a web mining technique, which has given solution to the above limitations. Its goal is to discover groups of objects (i.e., clusters), which are 'similar' between them and 'dissimilar' to the objects belonging to other clusters. Web clustering can involve either users (grouping of users who present similar browsing patterns) or pages (grouping of pages having related content) based on information derived from different sources. Specifically, user clustering approaches can be based on usage data (recorded in web server log files) and create groups of users with similar browsing behaviour (Petridou et al., 2008; Pallis et al., 2007). On the other hand, in web pages clustering approaches information can be extracted from pages content (denoted by keywords) (Hammouda and Kamel, 2004), structure (links between web pages or pages' structure as described by the involved tags) (Zhu et al., 2004) and usage data (which pages tend to be accessed by users with similar interests) (Su et al., 2002). Moreover, the clustering results may be beneficial for a wide range of applications such as websites' personalisation (Nasraoui et al., 2008), web caching and prefetching (Li et al., 2007), search engines (Hui et al., 2006) and Content Delivery Networks (Pallis and Vakali, 2006). In addition, the clustering results can contribute to the enhancement of recommendation engines (Chi et al., 2008) and to the design of collaborative filtering systems (Srinivasa and Medasani, 2004).

A clustering process needs to meet a number of challenges to be efficient. These challenges involve the definition of appropriate similarity or distance measures that will adequately capture the relations between data objects and guide properly the clustering process. The application of specific similarity (distance) measures depends on the underlying data nature and the data structures used for their representation (Jain et al., 1999). Next, effective grouping techniques need to be applied which will result in well-separated and compact clusters and which will handle, at the same time, particularities such as data high dimensionality and scalability in performance. Moreover, in every clustering process it is essential to analyse the obtained results by eliminating the irrelative deduction rules or patterns and extracting the interesting ones. The analysis of this data may provide valuable information on how to better structure a website and give an insight into the site's adequacy regarding the users' needs.

In this paper, we propose a clustering framework, which attempts to simultaneously cluster web users and pages, based on the users' access behaviour as recorded in a web server's log files. The proposed approach adopts a fuzzy in nature clustering scheme where users' clusters and consequently their members are related in a different degree with pages' clusters and their corresponding members These relations actually reveal the frequency (correlation) that the users of a cluster visit the set of pages that have also formed a cluster. More specifically, the proposed framework's main contributions can be summarised as follows:

- the proposed bi-clustering framework adapts spectral clustering theory principles, which proved to be quite efficient in data-clustering problems
- a different number of clusters may be identified by the proposed method for web users and pages
- the relations between web users' and pages' clusters are quantified to reveal users' interest in the different web pages' clusters.

The interpretation of the clustering results can be used towards improving the website's design, information availability and quality of provided services. The fuzzy bi-clustering nature of the proposed framework can significantly contribute towards this direction.

The remainder of the paper is organised as follows: Section 2 provides a brief overview of existing clustering approaches while in Section 3 our problem formulation is described. The proposed bi-clustering approach is given in Section 4 and the experimentation on both synthetic and real datasets is presented in Section 5. In Section 6, conclusions and future work insights are discussed.

2 Related work

Web clustering is a well-studied problem and numerous clustering algorithms appear in literature, which can be broadly categorised into different categories depending on the criteria employed. In a general categorisation scheme, clustering algorithms are divided into partitional and hierarchical, according to whether they produce flat partitions or a hierarchy of clusters (Jain et al., 1999; Xu and Wunsch, 2005). Another categorisation criterion of existing approaches refers to the data they involve (users or pages) or the nature of the grouping they perform and which may concern a hard assignment, i.e., data is divided into distinct clusters, where each data element belongs to exactly one cluster, or a fuzzy one, i.e., data elements are assigned to one or more clusters with different membership levels (Xu and Wunsch, 2005). Moreover, a clustering approach may be based on a distance function to identify the objects that should be clustered together (similarity-based) or to other probabilistic techniques (model-based) (Vakali et al., 2006).

More recent approaches extend the clustering problem of one dataset's objects (users or pages) and focus on the simultaneous grouping of both web pages and users by exploiting their relations (Zeng et al., 2002; Liu et al., 2005). The goal of these approaches is to identify groups of related web users and pages, which results from the tendency of some users to visit the same set of pages. This behaviour characterises users' interests as similar and highly related to the topic that the specific set of pages involves. The obtained results are particularly useful for applications such as e-commerce and recommendation engines, since relations between clients and products may be revealed. These relations are more meaningful than the one-way clustering of users or pages.

Furthermore, spectral clustering methods have attracted more and more attention given their promising performance in data clustering and simplicity in implementation. They treat the data-clustering problem as a graph-partitioning problem and they use information obtained from the eigenvalues and eigenvectors of the adjacency matrices

for the partitioning. Spectral clustering approaches have been successfully used in applications such as image segmentation (Shi and Malik, 2000) and social network analysis (Newman et al., 2002) and currently gain acceptance in the web domain.

Some indicative existing clustering approaches, which are based on usage data, are summarised in Table 1 and are categorised according to the aforementioned criteria.

Table 1 Clustering approaches

Approach	*Cluster content*	*Clustering algorithm*	*Methodology*
Cadez et al. (2002) and Pallis et al. (2007)	Web users	Partitional hard	Model based
Perkowitz and Etzioni (1998)	Web pages	Partitional hard	Similarity based
Petridou et al. (2008)	Web users	Partitional hard	Similarity based
Shokry et al. (2006) and Suryavanshi et al. (2005)	Web pages	Partitional fuzzy	Similarity based
Mojica et al. (2005)	Web pages	Hierarchical hard	Similarity based
Yang and Padmanabhan (2005)	Web users	Hierarchical hard	Similarity based
Lazzerini et al. (2003)	Web users	Hierarchical fuzzy	Similarity based
Castellano et al. (2007)	Web users	Partitional fuzzy	Similarity based
He et al. (2002) and Huang et al. (2006)	Web pages	Hard	Spectral clustering
Mobasher (1999) and Mobasher et al. (2002)	Web pages	Fuzzy	Similarity based
Zeng et al. (2002)	Web users/pages	Hard	Model based
Liu et al. (2005)	Web users/pages	Fuzzy	Similarity based

The clustering approach proposed in this paper aims to provide a framework for the simultaneous clustering of web pages and users in a way that relations between web users and pages will be adequately identified. A fuzzy scheme is developed where users grouped in the same users' cluster may be related to more than one web pages' cluster. Furthermore, the aforementioned relations are quantified to signify users' interest to the various web pages' clusters. This is quite important since, according to the authors' knowledge, existing approaches that apply simultaneous clustering on web users or pages provide either a hard assignment to clusters (Zeng et al., 2002) or a fuzzy one with no indication about the degree of relation between clusters (Liu et al., 2005). Moreover, the proposed framework employs spectral clustering theory principles which are simple to implement and can be solved efficiently by standard linear algebra methods (Luxburg, 2007). Results obtained by spectral clustering often outperform the traditional approaches employed in existing clustering frameworks.

3 Problem definition

We consider a particular web usage framework where we have (as a source) log files that capture the users' navigational behaviour. Let $U = \{u_1, \ldots, u_n\}$ denote the set of

n users and $P = \{p_1, \ldots, p_m\}$ be the set of m pages that the have been recorded in the log files.

Definition 1 (User visiting pattern)**:** Given a user u_i, $1 \le i \le n$, the user's visiting pattern is a multivariate vector consisting of m measurements $V(u_i,:) = (V(u_i, p_1), \ldots, V(u_i, p_m))$, where the element $V(u_i, p_j)$, $1 \le j \le m$, indicates the number of times the user u_i has visited the p_j page.

Example 1: Consider the vector $V(3, :) = (10, 31, 0, 44, 0)$, where $m = 5$. Then, the user identified as u_3 has 10, 31 and 44 visits to pages identified as 1, 2 and 4, respectively, but no visits to pages 3 and 5.

All the $V(u_i, :)$ are organised in the two-dimensional $u \times p$ users' pattern matrix V.

To obtain a more 'objective' perspective of the relations between users and pages, we proceed to the definition of the user's probability distribution.

Definition 2 (*User probability distribution*)**:** The probability distribution of the user u_i is a vector of m values produced by the normalisation of its $V(u_i, :)$ visiting pattern:

$$PV(u_i,:) = V(u_i,:) \Big/ \sum_{j=1}^{m} V(u_i, p_j). \tag{1}$$

The element $PV(u_i, p_j)$ expresses the probability with which the user u_i visits the p_j page.

Example 2: Considering the u_3 user of Example 1, its probability distribution vector results from $PV(3,:) = V(3,:) \Big/ \sum_{j=1}^{5} V(3, j)$ and thus $PV(3, :) = (0.12, 0.36, 0, 0.52, 0)$. Therefore, the user u_3 visits pages identified as 1, 2 and 4 with probabilities 0.12, 0.36 and 0.52, respectively whereas the probability to visit pages 3 and 5 is 0.

Our purpose is to create groups of related web users and pages. However, this relation cannot be defined in absolute terms, since a web user's interests cannot be limited to just one category of web pages. Thus, we aim to provide a solution to the following fuzzy correlation problem of web users and pages.

Problem 1 (Web users and pages fuzzy correlation)**:** *Given the set U of n users and P of m pages and their user probability distribution PV, find a set C of k subsets of users, a set C′ of k′ subsets of pages and a relation function f between pairs of the k and k′ subsets, such that*

$$\sum_{x=1}^{k} \sum_{y=1}^{k'} \sum_{u_i \in C_x, p_j \in C'_y} f(C_x, C'_y) PV(u_i, p_j) \text{ is maximised.}$$

The sum defined in the problem definition needs to be maximised because $f(C_x, C'_y)$ must be high for the clusters that involve users and pages connected with high visiting probabilities values ($PV(u_i, p_j)$).

Notation summary is given in Table 2.

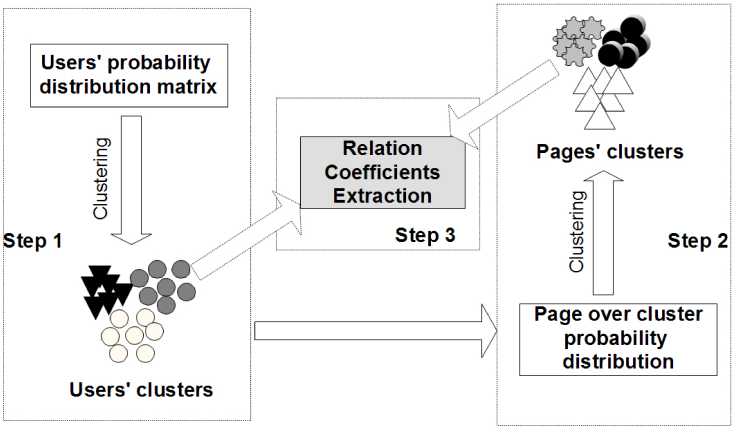
Table 2 Basic symbols notation

Symbol	*Description*
n	Number of users
m	Number of pages
U	Users' set $U = \{u_1, \ldots, u_n\}$
P	Pages' set $P = \{p_1, \ldots, p_m\}$
V	$n \times m$ users' pattern matrix
PV	$n \times m$ probability distribution matrix
f	Function of relations degrees

4 The fuzzy bi-clustering framework

According to Problem 1, our goal is to create k clusters of web users and k' clusters of web pages, which will be related in a different degree. The proposed bi-clustering approach is a three-step process, which initially creates the k users' clusters and then, based on them, proceeds to the extraction of k' web pages clusters. Considering the obtained clusters, relations between them are then extracted. The three phases of the overall process are clearly depicted in Figure 1.

Figure 1 The three-step bi-clustering process (see online version for colours)



In the first step of the clustering process, the users' access patterns, as recorded in the $u \times p$ probability distribution matrix PV, are used for the extraction of k users' clusters, with users having similar behaviour being clustered together. Then, the proposed approach proceeds to the second step, where the obtained users' clusters guide the pages clustering process. Therefore, it computes the relation (expressed in terms of visiting probabilities) between each users' cluster and pages to cluster the web pages, as indicated by the users' clustering results. The computed relations are organised on the page over cluster probability distribution matrix, which is then used for the extraction of the k' pages' clusters. In the third algorithm's step, the relations between the users' and pages clusters are computed.

4.1 Web users' clustering

In the first step of the proposed framework, the web users' clustering is performed based on the similarities of their visiting patterns, as recorded in their probability distribution vectors $PV(u_i, :)$. A common measure used to capture similarity between two (same dimension) vectors is the *Cosine Coefficient* (Zhang and Korfhage, 1999), which calculates the cosine of the angle between them.

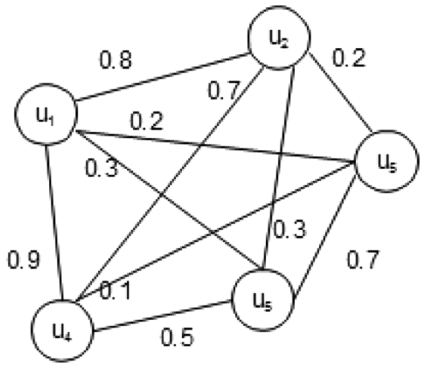
$$CS(u_i, u_j) = \frac{\mathrm{PV}(u_i, :) \cdot \mathrm{PV}(u_j, :)}{|\mathrm{PV}(u_i, :)| \cdot |\mathrm{PV}(u_j, :)|} = \frac{\sum_{l=1}^{m} PV(u_i, p_l) \cdot \mathrm{PV}(u_j, p_l)}{\sqrt{\sum_{l=1}^{m} PV(u_i, p_l)^2 \cdot \sum_{l=1}^{m} PV(u_j, p_l)^2}}. \quad (2)$$

The CS values fluctuate in the interval [0 … 1]. Values closer to 1 indicate higher similarity between the involved users. When the cosine similarity is 1 the two users are identical, when it is 0 users are unrelated.

Given the similarities between users we can use a graph structure to represent the users' dataset. Graphs are considered to be a convenient data structure for the representation of relations between sets of elements and have already been used in various clustering approaches (Xu et al., 2002; Dhillon, 2001).

Let us consider the weighted undirected graph $G = (U, E)$ presented in Figure 2, where $U = \{u_1, u_2, u_3, u_4, u_5\}$ the set of users and $E = \{\{u_i, u_j\} : u_i, u_j \in U, u_i \neq u_j\}$ the set of edges connecting users. The weight of each edge is equal to the similarity between the users that the edge connects.

Figure 2 Users graph representation



Creating k users' clusters means grouping together users with high similarity such that the quantity

$$\sum_{x=1}^{k} \sum_{u_i, u_j \in C_x} \mathrm{CS}(u_i, u_j) \text{ is maximised.}$$

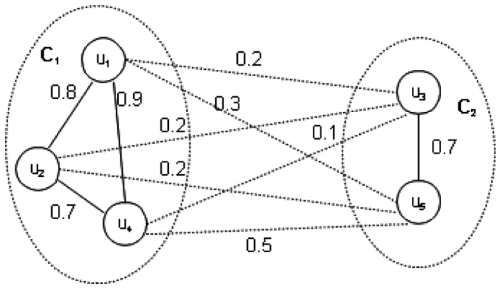
In case of the graph of Figure 2, its bi-partitioning, depicted in Figure 3, would result in the maximisation of the sum of similarities between the users belonging to the same cluster while, at the same time, the sum of similarities between the users of different clusters would be minimised. The last quantity corresponds to the graph cut minimisation which, in case of the two partitioning, is defined as:

$$\text{cut}(C_1, C_2) = \sum_{u_x \in C_1, u_y \in C_2} CS(u_x, u_y).$$

The above definition of the graph cut is easily extended to k subsets as follows (Dhillon, 2001):

$$\text{cut}(C_1, C_2, \ldots, C_k) = \sum_{i<j} \text{cut}(C_i, C_j).$$

Figure 3 Users graph cut



Hence, the two partitioning criteria, namely the minimisation of disassociation between the clusters and maximisation of the association can be satisfied simultaneously under the cut minimisation. Furthermore, the cut minimisation expresses the objective of the user clustering problem which is formulated in a way that

$$\sum_{u_i \in C_x} \sum_{u_j \in C_y} CS(u_i, u_j),$$

is minimised and $x, y = 1 \ldots k, x \neq y$.

Therefore, our problem is equal to the graph cut minimisation problem.

Moreover, Shi and Malik (2000) proposed the use of the normalised cut (*N*cut), which has proven to result in more balanced clusters and which is defined as:

$$N\text{cut}(C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} \frac{\text{cut}(C_i, \overline{C_i})}{\text{vol}(C_i)},$$

where $\overline{C_i}$ is the complement of C_i and $\text{vol}(C_i) = \sum_{u_x \in C_i, u_y \in U} CS(u_x, u_y)$ expresses the total relations from users in C_i to all users in the graph.

An approximate solution to the above normalised cut problem is obtained by following spectral clustering principles and solving the generalised eigenvalue system (Shi and Malik, 2000):

$$D^{-1/2}(D - CS)D^{-1/2}z = \lambda z \tag{3}$$

where D is the $n \times n$ diagonal degree matrix with

$$D(i,i) = \sum_{j=1}^{n} CS(u_i, u_j),$$

z a set of the eigenvectors of equation (3) and λ the corresponding eigenvalues. *Moreover*, $L = D^{-1/2}(D - CS)D^{-1/2}$ constitutes the normalised graph

Laplacian matrix (Luxburg, 2007). The k first eigenvectors of the normalised Laplacian matrix L provide the projection of the users similarities in the R^{uxk} space and are organised in increasing order (in terms of the respective eigenvalues) in the two-dimensional EV $(u \times k)$ table. Running the k-means algorithm on the EV table will result in the k-partitioning of the graph with minimum normalised cut and thus in the k users' clusters (Shi and Malik, 2000; Ng et al., 2002; Dhillon, 2001).

4.2 Web pages clustering

The second step in the proposed bi-clustering framework involves the clustering of web pages, which will be guided by the users' clusters obtained in the previous step. Given the $C_1, \ldots, C_k$ users clusters, we proceed to the definition of the $k \times p$ CP table as follows:

$$CP(C_x, p_y) = \frac{\sum_{u_i \in C_x} V(u_i, p_y)}{\sum_{u_i \in C_x} \sum_{j=1}^{m} V(u_i, p_j)}. \tag{4}$$

The element $CP(C_x, p_y)$ expresses the probability with which the users of cluster C_x visit the p_y page. Given the CP table of page over cluster probability distributions, we use the cosine similarity to calculate similarities between pages, in terms of the k clusters. Pages which are mostly visited by the same set of users' clusters are expected to present high values of similarity, while pages visited by different users' clusters presents low similarity values. The calculated values are organised in the two-dimensional $p \times p$ SP table. Thus, the $SP(p_i, p_j)$ value indicates the similarity between p_i and p_j pages in terms of the interest expressed by users' of the same cluster.

Following the procedure described in the web users' clustering step, we create a weighted undirected graph to represent similarities between the p pages. Then, the respective degree and Laplacian matrices are computed along with the k' eigenvectors of the Laplacian matrix. Running the k-means clustering algorithm on the eigenvectors' space will result on the $C'_1, \ldots, C'_{k'}$ web pages' clusters.

4.3 Relation coefficients

In the two steps described in the previous subsections, web users and pages are divided into k and k' clusters, respectively. The obtained clusters are related, since the web pages clusters were driven from the web users' clusters. Assigning a one-to-one relation between users and pages clusters is not proper, since web users are not strictly interested in one category of pages. Thus, our purpose is to provide a fuzzy in nature bi-clustering framework where, groups of related users are interested in a different degree to different groups of related web pages. Furthermore, this perspective offers to us the flexibility of resulting in a different number of users and pages clusters ($k \neq k'$).

The relation coefficients between the clusters are computed by the relation function $f(C_x, C'_y)$, which is defined as:

$$f(C_x, C'_y) = \frac{\sum_{u_i \in C_x} \sum_{p_j \in C_y} V(u_i, p_j)}{\sum_{u_i \in C_x} \sum_{l=1} V(u_i, p_l)}. \tag{5}$$

The relation coefficients indicate the frequency of the requested web pages that belong to a specific pages' cluster and are observed for each users' cluster. The sum of relation coefficients for each user cluster is equal to 1.

4.4 The bi-clustering algorithm

The proposed bi-clustering algorithm takes as input the users and pages sets U and V, respectively and produces as output the partitioning of users into the $C_1 \ldots C_k$ subsets, the partitioning of pages into the $C'_1,\ldots,C'_{k'}$ clusters and the relation coefficients $f(C_x, C'_y)$, which define the relation between the obtained users and pages clusters.

Algorithm 1: The fuzzy bi-clustering algorithm

Input: The set $U = \{u_1,\ldots,u_n\}$ of n users, the set $P = \{p_1,\ldots,p_m\}$ of m pages, and the k and k' numbers of clusters for users and pages, respectively.

Output: The sets $C = \{C_1,\ldots,C_k\}$ and $C' = \{C'_1,\ldots,C'_k\}$ of k users' clusters and k' pages clusters and the relation coefficients $f(C_x, C'_y)$

// users clustering

1: V = CreateVisitingPatterns (U, P);

2: PV = CreateProbabilityDistribution (V);

3: CS = FindUsersSimilarity (PV);

4: $D(i,i) = \sum_{j=1}^{n} CS(u_i, u_j)$

5: L = CalculateLaplacian (D, CS);

6: EV = EigenVectors (L, k);

7: C = k-means (EV, k);

// pages clustering

8: CP = Clusters2Pages (C, V);

9: SP = FindPagesSimilarity (CP);

10: $D'(i,i) = \sum_{j=1}^{m} SP(p_i, p_j)$

11: L′ = CalculateLaplacian (D′, SP);

12: EV′ = EigenVectors (L′, k′);

13: C′ = k-means (EV′, k′);

// relation coefficients

14: f(C_x, C_y) = CalculateRelationCoefficients (C, C′, V);

Given the U and P datasets, the algorithm computes the table of visiting patterns V (line 1), which describes the distribution of users' visits to the various pages. Then, it calculates the probability distribution matrix PV (line 2), which records the probability with which each user visits each page, according to equation (1). The PV matrix is then used for the calculation of similarities between users using the cosine coefficient (equation (2)). The obtained users' similarities are stored in the $u \times u$ CS table (line 3). Then, eigenvector analysis is performed on the CS table, which is also used for

the depiction of the users and their similarities in a graph format. First, the degree matrix D is computed (line 4) and then its corresponding normalised Laplacian (line 5). The k first eigenvectors of the Laplacian (line 6) are the input to the k-means algorithm, which results in the set C of k users clusters (line 7).

The second step of the algorithm performs the web pages clustering, which is guided by the k users' clusters. Initially, the CP table is created as defined in equation (4), which stores the probability with which users of each cluster visit the m pages (line 8). Then using the cosine coefficient (equation (2)), similarities between pages in terms of their relations with the k users clusters are computed and organised in the $p \times p$ table SP (line 9). The degree D' and Laplacian L' matrices of the SP table are calculated (lines 10, 11) and the Laplacian's k' eigenvectors, which form the EV' table (line 12), are used by k-means for the grouping of web pages into k' clusters (line 13).

In the final algorithm's step, the relation coefficients are computed using the relation function defined in equation (5) (line 14), to reveal the degree of relation between the k users and k' pages clusters.

5 Experimentation

To evaluate the proposed clustering approach, we carried out experimentation that involves both synthetic and real datasets. Both types of datasets were needed to capture the perspective of the proposed approach and its actual behaviour in real data workloads.

5.1 Data workload

Initially, we performed experimentation on synthetic data, which were generated in a way that users' clusters were in advance known and their access behaviour indicated their interests in web pages. The above experimentation was used to check whether the proposed method actually 'understands' and captures the underlying users' behaviour model that was originally used to generate the synthetic data, as well as their preferences in the various web pages, as denoted by their accessing behaviour.

In case of real datasets, users' navigational behaviour is recorded in web servers' log files. Our experimentation was based on log files that described users accessing behaviour in websites where a grouping of web pages would not be meaningless, since that could give us indications of how well our approach works. On the contrary, analysing users browsing behaviour over a website's pages that could not be distinguished into categories could not facilitate the clustering evaluation process. Furthermore, the datasets were chosen in a way that their size would not be prohibitive for a graphical representation of the clusters, which would contribute to a more comprehensible interpretation of the algorithm's results.

5.2 Clustering over synthetic datasets

We generated data based on a specific model and then checked whether the suggested method succeeded in discovering this model. More specifically, our synthetic datasets were generated as follows: we produced an $n \times m$ pattern table V whose data was divided in advance into k clusters (predefined clusters). For this $n \times m$ pattern table V, we fixed the dimensionality m of the data. Then, for each cluster we selected random

number of members while for each jth dimension ($j = 1, \ldots, m$), we selected a mean value $\mu_{i,j}$, which was uniformly distributed in [0, …, 99]. Points were then generated by adding a value sampled from the normal distribution $N(\mu_{i,j}, \sigma^2)$. The values for each of the selected mean value were at the same range for a set of pages to also identify k' sets of related pages.

To evaluate the algorithm's results, we proceeded to a graphical analysis approach, which is appropriate for representing multidimensional data. Graphical analysis is generally very important since it can reveal the underlying structure of a dataset. In case of high-dimensional data, advanced multivariate graphical techniques such as Andrews' curves are employed to efficiently depict the data properties (Andrews, 1972). Each multivariate observation (e.g., $(PV(u_i, p_1), \ldots, PV(u_i, p_m))$) where $i = 1, \ldots, n$, is transformed into a curve based on the function:

$$\begin{aligned} f(t) = {} & PV(u_i, p_1)\frac{1}{\sqrt{2}} + PV(u_i, p_2)\sin(t) \\ & + PV(u_i, p_3)\cos(t) + PV(u_i, p_4)\sin(2t) \\ & + PV(u_i, p_5)\cos(2t) + \ldots \end{aligned} \tag{6}$$

and plotted over the range $-\pi \le t \le \pi$. The definition of the $f(t)$ function is adjusted to the dimensionality of the plotted vector. Thus, each data point, in this case each user, may be viewed as a curve in the interval $[-\pi, \pi]$. These curves are representative of the dataset and can assist in distinguishing different groups, outliers, etc. Moreover, they have several interesting characteristics since they preserve the standard deviation and the distances of data points (e.g., close points will appear as close curves while distant points as distant curves). So, if there is an underlying structure in the data, it may be visible in its Andrews' curves. More specifically, regarding Andrews' curves in conjunction with clustering process, we can claim that the different shapes of curves among clusters are an indication of dissimilarity between users belonging to different clusters while the similar curves among the users of the same cluster are an indication of similarity between them (Theodosiou et al., 2008; Osorio et al., 2004; Spencer, 2003).

For this part of experimentation, we indicatively created a dataset of $n = 1000$ users, which belong to $k = 4$ clusters, while we fixed the number of pages at $m = 75$ and standard deviation $\sigma = 1$. The set of pages belong to five categories of 15 pages each of them This means that users that belong to the same cluster present similarities in their access behaviour with reference to the various sets of pages. Even though we experimented with the former values of n and m, as we have already discussed Andrews curves are scalable and appropriate for high-dimensional data.

Figure 4 provides a graphical representation of the $n = 1000$ users. Each user is represented by a single curve according to the definition of $f(t)$ function (equation (6)). The fact that there are different curves within the initial dataset proves that there exist differences in users' behaviour. Moreover, the existence of similar curves, which coincide, is evident and proves that there are users presenting similar visiting patterns.

After the clustering process, each of the four obtained users' clusters is depicted in each of the subplots of Figure 5. As we can see, curves have strong similarity within individual clusters while clusters' curves are separated and therefore well discriminated. Furthermore, we have four different curve shapes because users, in the synthetic dataset, were divided in advance into four clusters.

Figure 4 Users visiting patterns

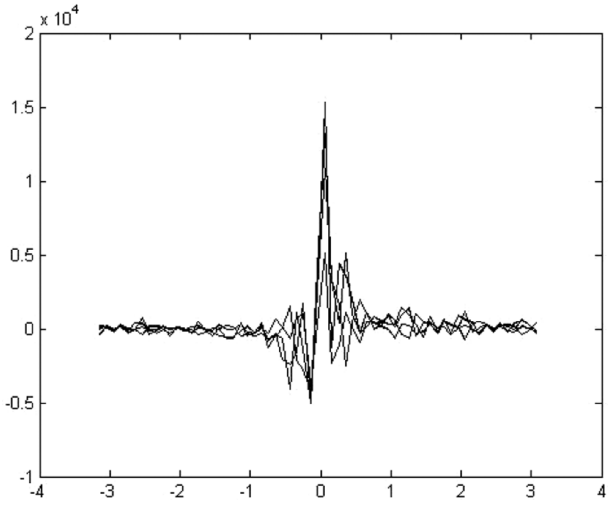
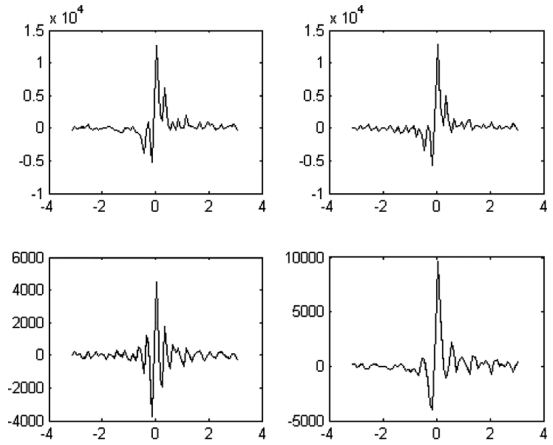
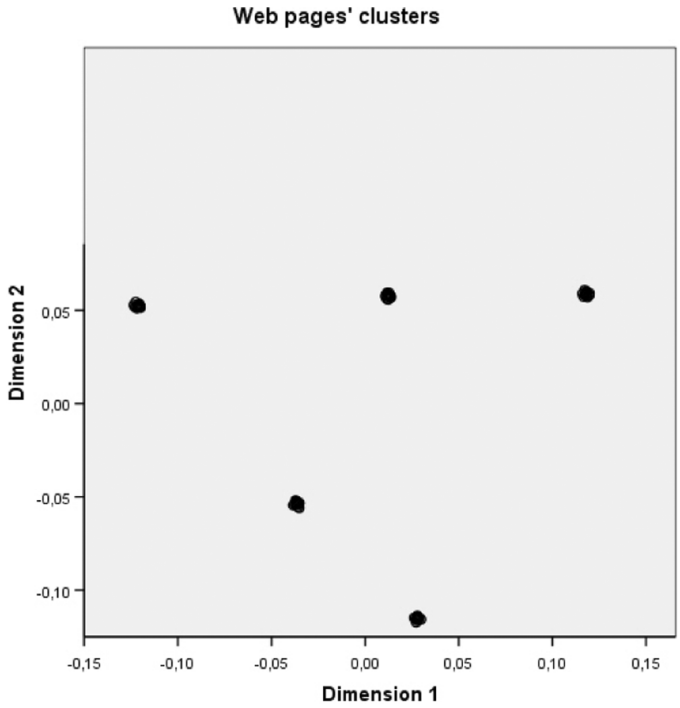


Figure 5 Users clusters



To visualise the five obtained pages' clusters, we have proceeded to the application of correspondence analysis method (Johnson and Wichern, 1998; Pallis et al., 2007). The goal of correspondence analysis is to describe the relationships between two categorical variables by projecting their values as points on a two-dimensional space, in such a way that the resulting plot describes the relationships between the categories of each variable. In our experimentation relations between pages are described using the correspondence analysis. As depicted in Figure 6, the algorithm manages to identify the five predefined pages clusters, which are characterised by a high degree of relation. Each of the obtained clusters contains 15 pages. These pages are grouped together because users who belong to same clusters show similar interest for these pages.

Figure 6 Pages' clusters



Users' interests are revealed from table of relation coefficients (Table 3), which shows which users and pages clusters are more related. For example, users of the first users' cluster (C_1) are mostly interested in pages contained in the fourth pages' cluster (C'_4) since the coefficient that corresponds to the C'_4 cluster has the highest value for the column that corresponds to the C_1 cluster. Similarly, users of the second users' cluster are more interested in pages contained in fifth pages cluster, etc.

Table 3 Relation coefficients

Clusters	C_1	C_2	C_3	C_4
C'_1	0.1811	0.2301	0.3786	0.1255
C'_2	0.1806	0.1550	0.0111	0.1261
C'_3	0.0137	0.0785	0.3034	0.3111
C'_4	0.4447	0.079	0.2287	0.0646
C'_5	0.1799	0.4570	0.0781	0.3726

Thus, the relation coefficients indicate the way in which users' clusters interest is distributed over all pages' clusters.

5.3 Experiments over real data workload

Our real data experimentation was based on two distinct sources of log files. The first source records users' navigational behaviour on an academic host (AUTH CSD department site[1]) while the second one logs users' visits on a general public popular website, which hosts pages about music machines.[2] In the CSD log file, the recorded entries referred to a six months period (October 2003–March 2004) and the music machines log file to a one month period (January, 1999).

Before using the two source files, we proceeded to their pre-processing, which is important and necessary in every knowledge extraction process. The pre-processing involves data cleaning, which removes any log entry that is not needed for the mining process (e.g., image files, css, swf or requests made by automated agents and spider programs). Thus, the initial log entries have been significantly reduced so as to work with useful information for the clustering. The remaining log entries of the CSD dataset involve 373 users and 255 pages, while for the music instruments dataset, 265 users and 127 pages.

We have run the bi-clustering algorithm for various numbers of users and pages clusters and we indicatively present the results for six users' and four pages' clusters in case of CSD and five users' and five pages' clusters for the music machines dataset. We chose the specific values for the numbers of clusters because, considering the design of the two websites, their pages could approximately be divided into the respective numbers of categories. The number of users' clusters was defined to be close to the number of pages, in order to find more 'absolute' relations between users and pages.

Figures 7 and 8 present the users' clusters of the CSD and music machines datasets, respectively, obtained at the end of the first step of the proposed clustering approach. The members of each cluster are denoted by the different markers while their distribution in space is in accordance with their between similarities, captured using the cosine coefficient (correspondence map).

Figure 7 Users' clusters of CSD dataset

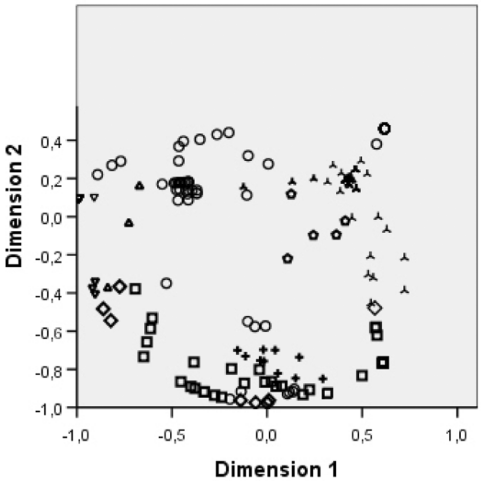
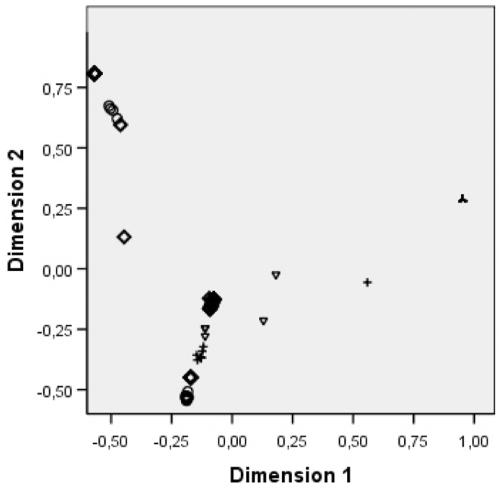


Figure 8 Users' clusters of music machines dataset



As depicted in both Figures 7 and 8, the clustering algorithm identifies groups of users, which are close in terms of the employed similarity measure. Moreover, it manages to reveal concave-shaped clusters, which is more evident in the case of the CSD dataset. Users of the music machines dataset present more similar visiting patterns and this is proved by their closeness and compactness of the formed clusters. Furthermore, there are users, who share common pages preferences and their data points on the correspondence map coincide.

Regarding the pages clusters it is more meaningful to proceed to their conceptual analysis since we are aware of their content. For both datasets, the pages that are clustered together do not all refer to the same topic areas. However, we have noticed that in each cluster there are pages that belong to a specific topic and appear more often than the others. Specifically, in case of the CSD dataset, we have identified the most often pages' topic for each of the four clusters: In the first cluster, pages that refer to studies and courses appear more often, while in the second cluster, we mostly meet home pages of the department's personnel. The third cluster contains mostly pages of the labs and studies and the fourth one refers to the index page and pages containing information about the post graduate studies.

On the other hand, the topics of the most often used pages of the music instruments dataset were harder to identify because the pages' contents are more restricted to information about manufacturers and instruments. Cluster 1 contains mostly pages that refer to information and searches about drum machines, Cluster 2 refers to pages that give general information about machines categories, Cluster 3 involves requests about music samples, Cluster 4 about sequencers and Cluster 5 particular manufacturers' (Casio, Roland) information, technical descriptions and images. Table 4 summarises the dominating topics per page cluster in both datasets.

Table 4 Pages' clusters content

CSD dataset	
C_1	Studies, courses
C_2	Personnel
C_3	Labs, courses
C_4	Index, information about postgraduate program
Music instruments dataset	
C_1	Drum machines
C_2	Information about machine categories
C_3	Music samples
C_4	Sequencers
C_5	Casio, Roland, technical descriptions, images

In the last section of our real data experimentation, we studied the relations between the obtained users and pages clusters. Figures 9 and 10 depict the aforementioned relations for the CSD and music machines datasets, respectively, based on the calculated relation coefficients.

As depicted in Figure 9, users' Clusters 1, 5 and 3 are closer to pages Cluster 1, which identifies them probably as students that are interested in courses web pages. Users' of Cluster 4 seem to be more interested in teachers' home pages (personnel), while users of the second cluster visit more the pages with information in post graduate studies. Thus, users of Cluster 4 may involve mostly teachers' or outside departments' visitors while users of the second cluster are probably post graduate students.

Similarly, in Figure 10, we can identify the relations between users and pages clusters of the music instruments dataset. Users of Cluster 5 are closely related to pages describing mostly sequencers while users of Cluster 1 are more interested in web pages about music samples. Furthermore, users of the third cluster visit more pages referring to drums and machine categories and users of Cluster 4 pages about specific manufacturers (Roland, Casio) as well as pages providing technical details and images of the requested instruments.

Figure 9 Clusters' relations for the CSD dataset

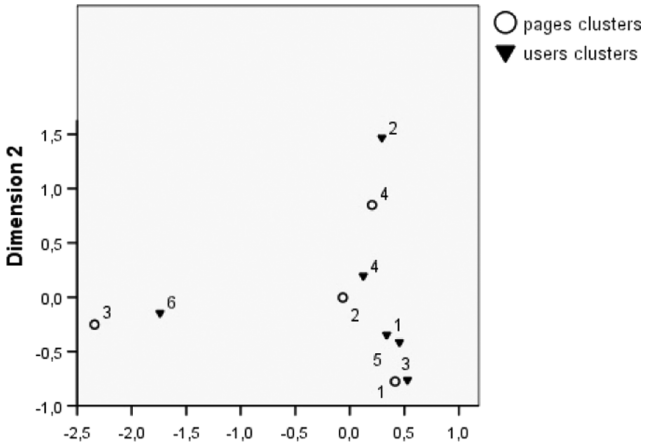
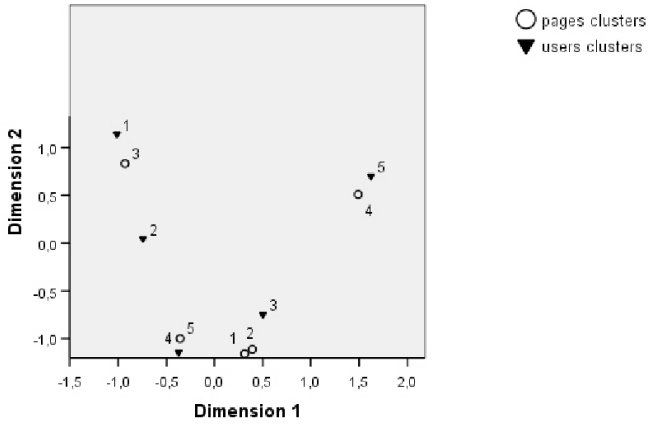


Figure 10 Clusters' relations for the music machines dataset



The analysis of the above results can be used by web administrators to reveal users' interests and offer more personalised information. Moreover, the graphical representation of the relations between clusters contributes to the identification of groups of users who tend to visit the same or different pages. These users may be indirectly related since they have similar but not exactly the same interests. Thus, the proposed fuzzy clustering scheme is advantageous, not only because it reveals close relations between clusters of users and pages, but also because it may identify indirectly related users or pages.

6 Conclusions and future work

This paper proposes a fuzzy bi-clustering approach, which aims to reveal relations between web users and web pages. The proposed framework is a three-step process which adopts the principles of spectral graph theory to provide an efficient and scalable solution based on the eigenvalues and eigenvectors of the adjacency matrix, for the clustering assignment. Initially, users' clusters are formed and then, these clusters guide the grouping of web pages. The degree of relations between the obtained clusters is also identified by the clustering process. Analysing clustering results could be beneficial in a great number of applications such as users' profiling for web personalisation systems and recommendation engines as well as efficient caching and prefetching policies. Moreover, the interpretation of the clustering results could facilitate acts of web administrators and designers towards improving web information retrieval, design and overall systems' performance.

Our future work aims at extending this framework by enriching the clustering process with information about web pages content. This could result in enhanced clusters, especially in terms of the pages' clusters, which could contribute in a clustering process of higher quality. Moreover, information about web pages' content would also make the two-directions clustering process meaningful. Specifically, pages' clusters would be initially defined based on their content and then these clusters would guide the users clustering process by also considering the way users access the pages of each cluster. Hence, a more enriched clustering process would result to enhanced clusters' quality and a more accurate definition of relation coefficients.

References

Andrews, D. (1972) 'Plots of high-dimensional data', *Biometrics*, Vol. 28, pp.125–136.

Cadez, I., Heckerman, D., Meek, C., Smyth, P. and White, S. (2002) *Visualization of Navigation Patterns on a Website Using Model-Based Clustering*, Tech. Report MSR-TR-00-18, Microsoft Research.

Castellano, G., Fanelli, A.M., Mencar, C. and Torsello, M.A. (2007) 'Similarity-based fuzzy clustering for user profiling', *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops*, November, Silicon Valley, pp.75–78.

Chi, C-C., Kuo, C-H., Lu, M-Y. and Tsao, N-L. (2008) 'Concept-based pages recommendation by using cluster algorithm', *Proceedings of Eighth IEEE International Conference on Advanced Learning Technologies*, 1–5 July, Santander, Spain, pp.298–300.

Dhillon, I.S. (2001) 'Co-clustering documents and words using bipartite spectral graph partitioning, *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 26–29 August, San Francisco, California, pp.269–274.

Eirinaki, M. and Vazirgiannis, M. (2003) 'Web mining for web personalization', *ACM Transactions on Internet Technology*, Vol. 3, No. 1, pp.1–27.

Hammouda, K.M and Kamel, M.S. (2004) 'Efficient phrase-based document indexing for Web document clustering', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 10, pp.1279–1296.

He, X., Zha, H., Ding, C. and Simon, H. (2002) 'Web document clustering using hyperlink structures', *Computational Statistics and Data Analysis*, Vol. 41, No. 1, pp.19–45.

Huang,, J., Zhu, T. and Schuurmans, D. (2006) 'Web communities identification from random walks', *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, September, Berlin, Germany, pp.187–198.

Hui, Z., Bin, P., Ke, X. and Hui, W. (2006) 'An efficient algorithm for clustering search engine results', *2006 International Conference on Computational Intelligence and Security*, November, Guangzhou, China, pp.1429–1434.

Jain, A., Murty, M. and Flynn, P. (1999) 'Data clustering: a review', *ACM Computing Surveys*, Vol. 31, No. 3, pp.264–323.

Johnson, R.A. and Wichern, D.W. (1998) *Applied Multivariate Statistical Analysis*, Prentice-Hall, Upper Saddle River.

Lazzerini, B., Marcelloni, F. and Cococcioni, M. (2003) 'A system based on hierarchical fuzzy clustering for web users profiling', *IEEE International Conference on Systems, Man and Cybernetics*, October, Asian Liaison, pp.1995–2000.

Li, H-Y., Xie, C-S. and Liu, Y. (2007) 'A new method of pefetching I/O requests', *2007 International Conference on Networking, Architecture and Storage*, July, Guilin, China, pp.217–224.

Liu, B. (2007) *Web Data Mining Exploring Hyperlinks, Contents and Usage Data*, 1st ed., Book 1st Chapter, Springer.

Liu, X., He, P. and Yang, Q. (2005) 'Mining user access patterns based on Web logs', *Canadian Conference on Electrical and Computer Engineering*, May, Saskatoon Inn Saskatoon, Saskatchewan Canada, pp.2280–2283.

Luxburg, U. (2007) 'A tutorial on spectral clustering', *Statistics and Computing*, Vol. 17, No. 4, pp.395–416.

Mobasher, B. (1999) 'A web personalization engine based on user transaction clustering', *Proceedings of the 9th Workshop on Information Technologies and Systems*, December, Charlotte, North Carolina, pp.179–184.

Mobasher, B., Honghua, D., Luo, T. and Nakagawa, M. (2002) 'Discovery and evaluation of aggregate usage profiles for web personalization', *Data Mining and Knowledge Discovery*, Vol. 6, No. 1, pp.61–82.

Mojica, J.A., Rojas, D.A., Gomez, J. and Gonzalez, F. (2005) 'Page clustering using a distance based algorithm', *Third Latin American Web Congress*, October, Buenos Aires, Argentina, p.7.

Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R. (2008) 'A web usage mining framework for mining evolving user profiles in dynamic websites', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 2, pp.202–215.

Newman, M., Watts, D. and Strogatz, S. (2002) 'Random graph models of social networks', *Proceedings of the National Academy of Sciences USA*, pp.2566–2572.

Ng, A.Y., Jordan, M.I. and Weiss, Y. (2002) 'On spectral clustering: analysis and an algorithm', in Dietterich, T., Becker, S. and Ghahramani, Z. (Eds.): *Advances in Neural Information Processing Systems*, MIT Press, p.14.

Osorio, C-G., Maudes, J. and Fyfe, C. (2004) 'Using Andrews curves for clustering and sub-clustering self-organizing maps', *European Symposium on Artificial Neural Networks*, April, Bruges, Belgium, pp.477–482.

Pallis, G. and Vakali, A. (2006) 'Insight and perspectives for content delivery networks'. *Communications of the ACM*, Vol. 49, No. 1, pp.101–106.

Pallis, G., Angelis, L. and Vakali, A. (2007) 'Validation and interpretation of web users' sessions clusters. Information', *Processing and Management*, Vol. 43, No. 5, pp.1348–1367.

Perkowitz, M. and Etzioni, O. (1998) 'Adaptive web sites: automatically synthesizing web pages', *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, July, Madison, Wisconsin, pp.727–732.

Petridou, S., Koutsonikola, V., Vakali, A. and Papadimitriou, G. (2008) 'Time-aware web users clustering', *IEEE Transaction on Knowledge and Data Engineering*, Vol. 20, No. 5, pp.653–667.

Shi, J. and Malik, J. (2000) 'Normalized cuts and image segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp.888–905.

Shokry, R.A., Saad, A.A., El-Makkey, N.M. and lsmail, M.A. (2006) 'Using new soft clustering technique in adaptive web site', *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, December, Hong Kong, pp.281–286.

Spencer, N. (2003) 'Investigating data with Andrews plots', *Social Science Computer Review*, Vol. 21, No. 2, pp.244–249.

Srinivasa, N. and Medasani, S. (2004) 'Active fuzzy clustering for collaborative filtering', *Proceedings of IEEE International Conference on Fuzzy Systems*, July, Budapest, Hungary, pp.1607–1702.

Su, Z., Yang, Q., Zhang, H., Xu, X., Hu, Y-H. and Ma, S. (2002) 'Correlation-based web document clustering for adaptive web interface design', *Knowledge and Information Systems*, Vol. 4, No. 2, pp.151–167.

Suryavanshi, B., Shiri, N. and Mudur, S. (2005) 'An efficient technique for mining usage profiles using relational fuzzy subtractive clustering', *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration*, April, Tokyo, Japan, pp.23–29.

Theodosiou, T., Angelis, L., Vakali, A. and Thomopoulos, G. (2008) 'Gene functional annotation by statistical analysis of biomedical articles', *International Journal of Medical Informatics*, Vol. 76, No. 8, pp.601–613.

Vakali, A., Pallis, G. and Angelis, L. (2006) 'Clustering web information sources, published in the book', in Vakali, A. and Pallis, G. (Eds.): *Web Data Management Practices: Emerging Techniques and Technologies*, Idea-Group Publishing, USA, pp.34–55.

Xu, R. and Wunsch, D.I. (2005) 'Survey of clustering algorithms', *IEEE Transactions on Neural Networks*, Vol. 16, No. 3, pp.645–678.

Xu, Y., Olman, V. and Xu, D. (2002) 'Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning tree', *Bioinformatics*, Vol. 18, No. 4, pp.526–535.

Yang, Y. and Padmanabhan, B. (2005) 'GHIC: a hierarchical pattern-based clustering algorithm for grouping web transactions', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 9, pp.1300–1304.

Zeng, H-J., Chen, Z. and Ma, Y-M. (2002) 'A unified framework for clustering heterogeneous web objects', *Proceedings of the 3rd International Conference on Web Information Systems Engineering*, December, Singapore, pp.161–172.

Zhang, J. and Korfhage, R. (1999) 'A distance and angle similarity measure method', *Journal of the American Society for Information Science*, Vol. 50, pp.772–778.

Zhu, J., Hong, J. and Hughes, J. (2004) 'PageCluster: mining conceptual link hierarchies from web log files for adaptive web site navigation', *ACM Transactions on Internet Technology*, Vol. 4, No. 2, pp.185–208.

Notes

[1]Department of Informatics AUTH, http://www.csd.auth.gr

[2]Web logs for the music machines, http://www.cs.washington.edu/research/adaptive/download.html