A Divergence-Oriented Approach for Web Users Clustering

Sophia G. Petridou, Vassiliki A. Koutsonikola,
Athena I. Vakali, and Georgios I. Papadimitriou

Dept of Informatics Aristotle University,
54124 Thessaloniki, Greece
{spetrido, vkoutson, avakali, gp}@csd.auth.gr

Abstract. Clustering web users based on their access patterns is a quite significant task in Web Usage Mining. Further to clustering it is important to evaluate the resulted clusters in order to choose the best clustering for a particular framework. This paper examines the usage of Kullback-Leibler divergence, an information theoretic distance, in conjuction with the k-means clustering algorithm. It compares KL-divergence with other well known distance measures (Euclidean, Standardized Euclidean and Manhattan) and evaluates clustering results using both objective function's value and Davies-Bouldin index. Since it is imperative to assess whether the results of a clustering process are susceptible to noise, especially in noisy environments such as Web environment, our approach takes the impact of noise into account. The clusters obtained with KL approach seem to be superior to those obtained with the other distance measures in case our data have been corrupted by noise.

1 Introduction

Web Data Mining, the application of data mining techniques on Web data to obtain knowledge, has become an important research area since the amount of information on the Web is increasing at tremendously fast rates. According to [1, 2] Web data can be handled as *usage, content*, *structure*, or *user profile* and collected from different sources such as *server*, *client* or *proxy* log files. Here we focus on server side usage data as we are interested in Web Usage Mining. Server logs keep information about multiple users who access a single site and provide usage data which include IP addresses, url requests, responding codes and the date and time of accesses according to the type of logging level defined in the Web server configuration file. However, the collected data might not be entirely reliable as the cached page requests [3] are not recorded in log file or, on the other hand, the search engines or virus created requests are logged without being necessary for mining process. As the quality of data that will be processed is a key issue for data mining in general, data preprocessing is important so as to increase mining's accuracy and reliability.

After preprocessing, the data clustering is taking place. Clustering creates groups of items (clusters) which are "similar" between them and "dissimilar" to the items belonging to other clusters. Web Usage Mining clustering can involve either users or pages. The purpose of user clustering is to establish groups of users that present similar browsing patterns while page clustering discovers groups of pages having related

M. Gavrilova et al. (Eds.): ICCSA 2006, LNCS 3981, pp. 1229 – 1238, 2006.

content. Typical usage clustering techniques are divided into partitional and hierarchical [4]. Partitional clustering algorithms attempt to create a specific number of clusters that optimize a criterion function while hierarchical builds (agglomerative) or breaks up (divisive) a hierarchy of clusters. A clustering algorithm is characterized by the proximity measure that quantifies how "similar" two data points are. For example, k-means [5] is a commonly used partitional clustering algorithm that attempts to minimize an objective function value which measures the distance of each point from the center of the cluster to which the point belongs.

But, whatever clustering algorithm is chosen, it is imperative to assess whether the results are susceptible to noise [6]. In the Web environment, "noise" refers to visits which are executed by chance, by mistake or with remote probability. In general, it would be wishful a user clustering process not to be seriously affected by these random events. Here, we assess the results of unsupervised clustering with the KL-divergence from information theory being compared as an alternative to the more commonly used distance measures such as Euclidean, Standardized Euclidean or Manhattan distance. The evaluation process is accomplished using various validity indices [7, 8]. This paper evaluates clusters using both objective function's value and Davies-Bouldin index and takes into account the impact of noise. The clusters obtained with the KL-divergence approach were found to be better to those obtained using the traditional distances in the presence of noise. The KL-divergence has already been used in cluster analysis of biological data [9] and also led to superior patterns compared to those that a hierarchical clustering algorithm produced using the Pearson correlation distance measure.

A word about notation: upper-case letters such as X, Y will denote random variables while lower-case letters such as x, y individual set elements. Probability distributions will be denoted by p, q when the random variable is obvious or by p(X) to make the random variable explicit. Different clusters will be donated as C_i, C_j, etc whereas the center of a cluster as c_i, c_j etc.

The remainder of this paper is organized as follows. Section 2 presents our divergence-oriented clustering approach and is divided in Section 2.1 which gives emphasis on the representation of Web data and compares the KL-divergence with other distance measures, Section 2.2 where we present our clustering algorithm and Section 2.3 where we discuss the clustering evaluation. Section 3 exhibits our implementation and experimental results that show the superiority of our clustering process in noisy data. Finally, we summarize our conclusions and discuss future work in Section 4.

2 Divergence-Oriented Clustering

The basic steps of our clustering approach are presented in Figure 1 and can be summarised as follows:

- *Data preprocessing*: the collected data are processed so as we get rid of meaningless information (i.e. search engines requests). As a result, we create a table each row of which represents a user and each column corresponds to a page of the web site. Each cell of this table indicates how many times a user visits a page.
- *Clustering algorithm*: in our clustering process the data are first normalized and then classified using the k-means algorithm with the KL-divergence. Each cell of

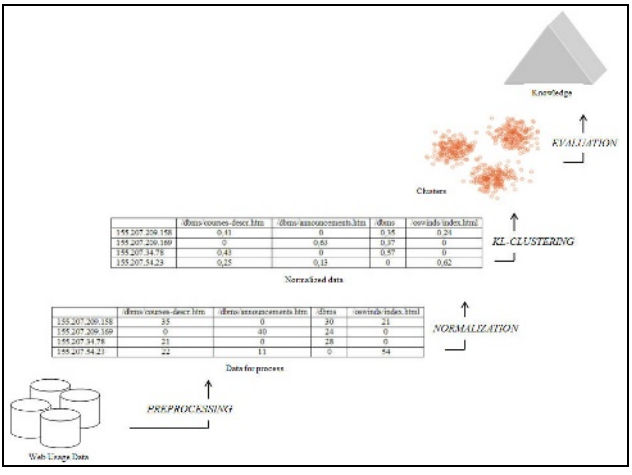


Fig. 1. Steps in the clustering process

our normalized table expresses the probability with which a user will visit a page and each row is the probability distribution of each user (p, q). This table as well as the number of clusters to be created is the input to our algorithm.

- *Clustering Evaluation*: created clusters are evaluated in order to assess the quality of our approach and extract knowledge.

2.1 Measuring Web Usage Data Distances

Clustering approaches identify objects that are simiral to each other by using certain distances such as the Euclidean, Standardized Euclidean and Manhattan distance [10]. Given the probability distributions p, q and r these distances satisfy the following five properties of a metric:

- Non-negativity $d(p,q) >= 0$
- Definiteness $d(p,q)=0$ iff $p(x)=q(x)$
- Identification mark $d(p,p)=0$
- Symmetry $d(p,q)=d(q,p)$
- Triangle inequality $d(p,q)>=d(p,r)+d(r,q)$

At the same time, information theory methods, such as KL-divergence, proved to be suited for the clustering of biological data [9] as are capable of assessing similarities and dissimilarities between data distributions. A KL-divergence approach in biological data clustering [9] led to superior patterns compared to those that a hierarchical clustering algorithm produced using the Pearson correlation distance measure.

The relative entropy or Kullback-Leibler divergence, which originated from information theory, is a measure of the "distance" between two probability distributions but it is not a true metric since it is not symmetric and does not obey the triangle inequality. Given two probability distributions p(x) and q(x) of a discrete variable the KL-divergence is a quantity which measures the difference between p(x) and q(x) and is defined as follows [11, 12]:

$$KL(p,q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (1)$$

Although there are a number of common features that characterize both web and biological data, the KL-divergence has not been explored for Web data analysis applications. Here we propose using KL-divergence in web clustering since web users' data and biological data are both characterized by:

- *Huge exploration spaces*
- *Complexities and dynamic data nature*
- *Fast searching and retrieval rates*
- *Data representation*

Here, we assess the performance of unsupervised clustering using the KL-divergence in Web data and we compare KL-divergence with the commonly used Euclidean, Standardized Euclidean and Manhattan distances. Our intention is to show that due to the fact that KL-divergence is a probabilistic distance measure instead of an actual distance it is less susceptible to noisy environments such as the Web and so it leads to better results in the presence of noise.

2.2 The KL-Divergence Clustering Approach

Our approach is a two-step process where the data first normalized and then classified using the k-means algorithm with the KL-divergence as the dissimilarity measure used for clustering.

Data normalization
As it is shown in Figure 1, after the data preprocessing, we create a table (nxm) where each row corresponds to a user, each column to a page and each table cell indicates how many times a user visits a page. This table is normalized in order to produce a table (nxm) where its elements are the probabilities with which each user visits each page. The normalized expression values for each user fall in the interval [0, 1] and each row sum is 1 (unit total probability mass). Each row of this second table is the probability distribution of each user and is suitable for the calculation of distances between the users. After this calculation we receive the distance table (nxn) which is symmetric and will be the input to the clustering algorithm with the number of clusters to be created. In the KL distance table the divergence between p and q is defined as KL(p,q)+KL(q,p) which is symmetric and nonnegative [11, 12].

K-means clustering method
The k-means is an unsupervised, partitional learning algorithm which classifies a given data set to a certain number of clusters (assume k clusters) fixed a priori. It begins by initializing a set of k centres, one for each cluster. These centres should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. Then it assigns each object of the data set to the cluster whose centre is the nearest. When no point is pending, the first step is completed and an early clustering is done. At this

point we need to re-compute k new centres. After we have these k new centres, a new binding has to be done between the same data set points and the nearest new centre. A loop has been generated. As a result of this loop we may notice that the k centres change their location until no more changes are done. In other words centres stop changing. Finally, this algorithm aims at minimizing an *objective function (J)*, in this case a squared error function. The objective function:

$$J(X) = \sum_{j=1}^{k} \sum_{i=1}^{n} d(x_i^{(j)} - c_j) \tag{2}$$

where $d(x_i^{(j)} - c_j)$ is the chosen distance measure (Euclidean, Standardized Euclidean, Manhattan or KL-divergence) between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the data points from their perspective cluster centres.

The k-means algorithm is composed of the following steps:

```
Select k points as the initial centres: c1, c2, …, ck
Assign all points to the group that has the closest
centre
Re-compute the centre of each cluster
Repeat steps 2 and 3 until the centres don't change or
when the objective function improvement between two
consecutive iteration is less than a minimum amount of
improvement specified
```

2.3 Clustering Evaluation

Since various clustering algorithms result in different clusters it is important to perform an evaluation of the results to assess their quality. In clustering, the procedure of evaluating the results is known as cluster validation and can be based on various measures called validity measures.

The validity measures are divided in two categories depending on whether they have any reference to external knowledge [13]. By external knowledge we refer to a pre-specified structure which reflects our intuition about the clustering structure of a data set. The measures that have no reference to external knowledge are called internal quality measures and they are estimated in terms of quantities that involve the data set.

Dunn's index [14] and DB index [15] are two internal quality measures that have a close relationship in that they both try to minimise the within-cluster scatter while maximising the between-cluster separation in order to find compact and well separated clusters. DB index is more robust than Dunn's index.

A broadly accepted and quite reliable external measure is the F-Measure, which combines the precision and recall ideas from information retrieval [16].

In our approach we use DB index to perform clustering validation. Given that K is the number of clusters, C_i and C_j are the closest clusters according to average distance d and diam is the diameter of a cluster, the DB index is defined as follows:

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left[\frac{diam(C_i) + diam(C_j)}{d(C_i, C_j)} \right] \tag{3}$$

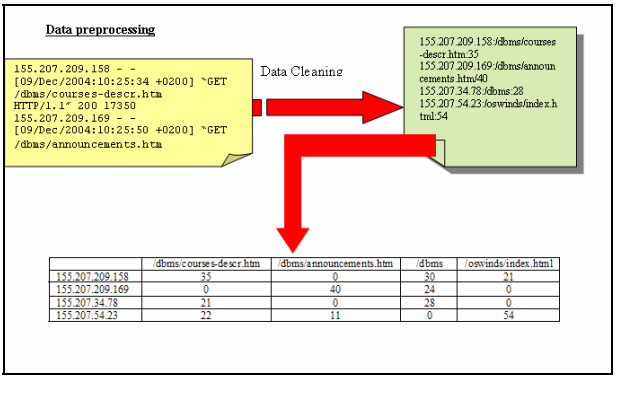
3 Implementation

3.1 Data Workload

In our usage mining approach, data is collected from user's interaction with the web, which is recorded in web server's log file. Our experiments were conducted based on the user's behavior that was extracted by the log files of AUTH Computer Science Department web server. The log files that were used contained log entries for a period of two months.

Data preprocessing presented in Figure 2 involved data cleaning, the process that removes any log entry that is not needed for the mining process. Typically, these entries refer to image files, css, swf as well as recorded requests by ip addresses that do not behave normally. In addition, log entries with status other than 200 which indicates success and 304 which indicates redirection, are being removed. Furthermore according to [17] data cleaning involves removing log entries that are either negligible to influence the results or could dominate the clustering process. In our experiments, we exclude from the clustering process users that have less than 5 visits because they are considered not to influence the clustering process. We also exclude users that have more than 280 because they are too many visits compared to the number of other users' visits and that could possibly mislead the clustering process.

Data preprocessing results in a table each cell of which indicates the number of times a user visits a page.



	/dbms/courses-descr.htm	/dbms/announcements.htm	/dbms	/oswinds/index.html
155.207.209.158	35	0	30	21
155.207.209.169	0	40	24	0
155.207.34.78	21	0	28	0
155.207.54.23	22	11	0	54

Fig. 2. Data Preprocessing

3.2 Clustering Using Various Distances

In our clustering approach, we examined the clusters obtained with the KL divergence and three others distance measures, namely Euclidean, Standardized Euclidean and Manhattan. For each distance measure, the number of clusters varied from 3 to 10 (fixed a priori) and the results shown are over an average of 1000 runs.

Almost all of the clusters were well populated, with only a few clusters with single users. Figure 3 presents the objective function in case of algorithm handling the actual

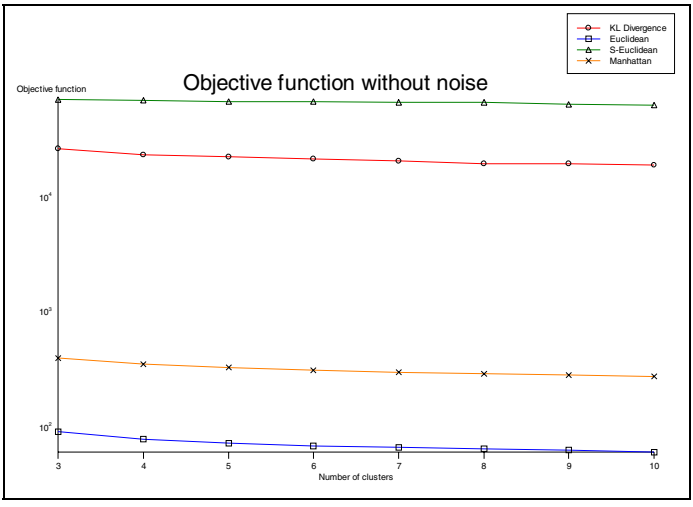


Fig. 3. Objective function values without noise

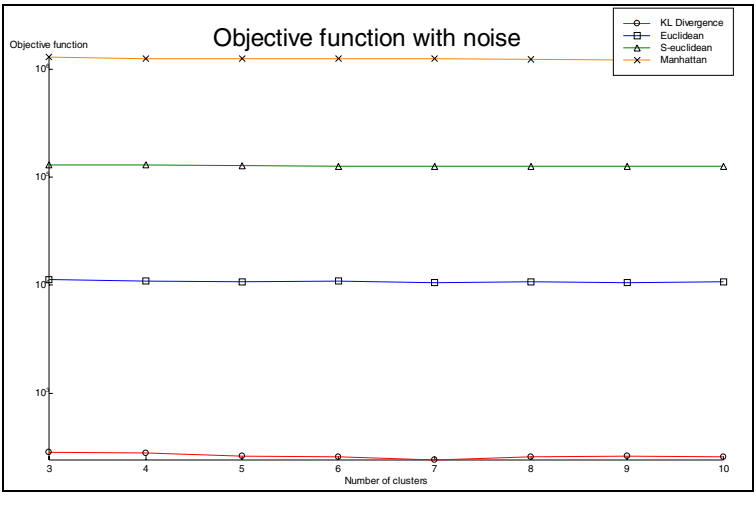


Fig. 4. Objective function values with noise

distance table for the different distance measures while Figure 4 presents the objective function when table values have slightly been altered using a random function. This change assimilates the noisy web data which would refer to accidental or rare visits.

Figure 3 indicates that performance of KL-divergence approach is superior to that using Standardized Euclidean distance but clustering using the Euclidean and Manhattan gave better results.

In figure 4 we can observe that clustering based on KL-divergence minimizes the value of objective function compared to other distances when we randomly add noise in our distance table. What is more, there is a decrease in the values of objective

function in the case of KL-divergence in the presence of noise while when we use the other distances the values are increased. As the objective function is an indicator of the distance of the data points from their perspective cluster centres this means that KL clustering gives more cohesive clusters without, however, damaging seriously the quality of clusters as it will be shown by DB index values. We conclude that Euclidean, Standardized Euclidean and Manhattan distances are more sensitive to noise when compared to the KL-divergence and the explanation of this is based on the fact that the three distances are true distances while the KL-divergence is a probabilistic measure of distance.

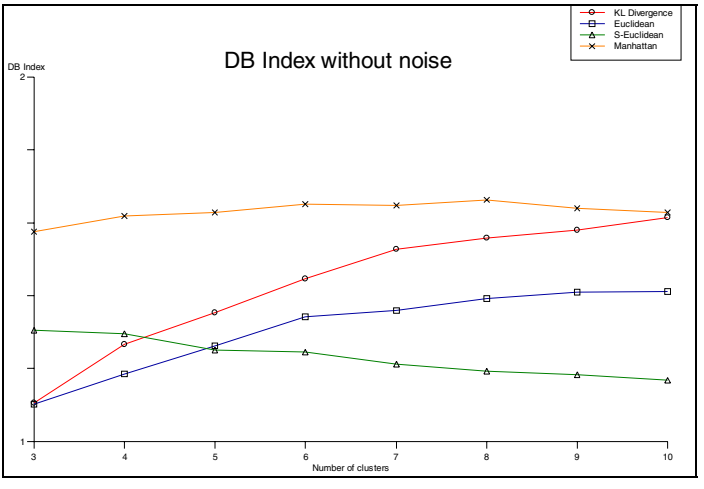


Fig. 5. The DB index validity without noise

In the second section of our experiments we used Davies-Bouldin index to evaluate clustering results. Figure 5 represents DB Index values when there is no noise and indicates that performance of KL-divergence approach is comparable to approaches based on true distances measures and it particularly outperforms Manhattan approach for the different values of k. In general, we conclude that our approach behaves in a similar way with the other real distance approaches in case of noise absence meaning that db index values change in a quite regular pattern.

With the addition of noise our approach behaves better than the others as it is shown in Figure 6. They are all improving as k increases but the curve of KL-divergence approach has steeper gradient and gives lower values of DB Index. This behaviour can be explained considering that when adding noise to our data set the real distance between data points measured by Euclidean, S-Euclidean and Manhattan distances is increased and the clustering process results in less solid clusters. On the other hand, the KL-divergence as a probabilistic measure is not affected by the presence of noise as much as the other distances and gives better clustering results.

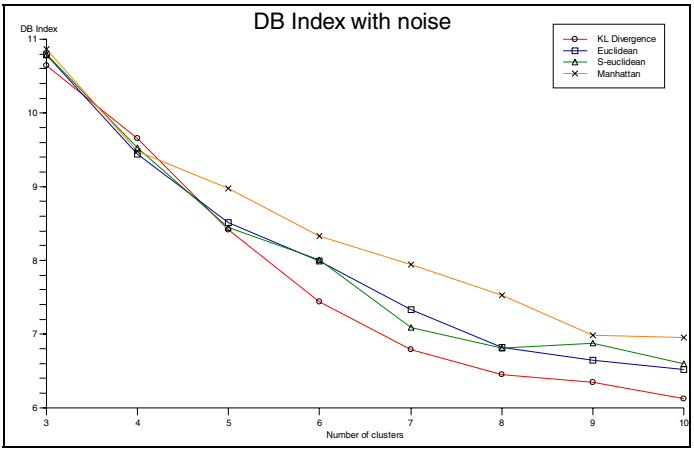


Fig. 6. The DB index validity with noise

4 Conclusion and Future Work

In this paper we introduced the usage of KL-divergence, a probabilistic measure which has already been used in mining biological data, in the area of web users clustering. In addition, we used the objective function values and DB-index validity measure so as to evaluate the results of our experiments. The results indicated that our approach behaves satisfactorily compared to other real distance based approaches and overcomes them with the presence of noise which is important as the web environment is a noisy one. Our next step is to experiment with other validity measures and compare our KL-divergence approach with other partitional and hierarchical algorithms.

References

1. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Exploratios, Vol.1, No. 2, Jan 2000.
2. S. Petridou, G. Pallis, A. Vakali, G. Papadimitriou, A. Pomportsis: Web Data Accessing and the Web Searching Process, ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'03), Tunis, Tunisia, July 14-18, 2003.
3. A.Vakali and G. Papadimitriou: Web Engineering: The Evolution of New Technologies, Guest Editorial in IEEE Computing in Science and Engineering, 6(4): 10-11, Jul./Aug. 2004.
4. Anil K. Jain and Richard C. Dubes: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey, 1988.
5. J.B. McQueen: Some Methods for Classification and Analysis of Multivariate Observations. Proc. 5th Berkley Symposium on Mathematical Statistics and Probability, I: Statistics, 1994, pp, 281-297.

6. Kerr MK and Churchill GA: Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments, PNAS, 98:8961-8965, 2001.
7. Benno Stein, Sven Meyer zu Eissen, Frank Wißbrock: On Cluster Validity and the Information Need of Users, 3rd IASTED Int. Conference on Artificial Intelligence and Applications (AIA 03).
8. Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis: Clustering Validity Checking Methods: Part II, , SIGMOD Record, Vol. 31, No. 3, September 2002.
9. Jyotsna Kasturi 1 , Raj Acharya1 and Murali Ramanathan2: An information theoretic approach for analyzing temporal patterns of gene expression. Bioinformatics 19, 4, 2003, 449--458.
10. A. Sturn: Cluster analysis for large scale gene expression studies. Master's thesis, Graz University of Technology, Graz, Austria, 2001
11. Inderjit S. Dhillon, Subramanyam Mallela and Rahul Kumar: Enchanced Word Clustering for Hierarchical Text Classification, KDD 2002: 191-200.
12. Inderjit S. Dhillon, Subramanyam Mallela and Rahul Kumar: Information Theoretic Feature Clustering for Text Classification, Journal of Machine Learning Research 3: 1265-1287 (2003).
13. Francois Boutin, Mountaz Hascoer: Cluster Validity Indices for Graph Partitioning, Proceedings of the Eighth International Conference on Information Visualisation (IV'04) 1093-9547/04 IEEE.
14. J.C. Dunn:Well separated clusters and optimal fuzzy partitions. J. Cybern., vol. 4, no. 3, pp. 95-104, 1974.
15. D.L. Davies and D.W. Bouldin: A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Learning, 1(2), 1979.
16. B. Larsen and Ch. Aone. Fast and Effective: Text Mining Using Linear-time Document Clustering. In Proc. KDD 99 Workshop San Diego USA, San Diego, CA, USA, 1999.
17. Bamshad Mobasher, Robert Cooley, Jaideep Srivastava: Creating Adaptive Web Sites Through Usage-Based Clustering of URLs, In Proccedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99).