IMPROVEMENT OF A METHOD BASED ON HIDDEN MARKOV MODEL FOR CLUSTERING WEB USERS

Sadegh Khanpour[1] and Omid sojoodi[2]

[1]Faculty of Electrical, Computer and IT Engineering,
Qazvin Azad University, Qazvin, Iran
sadeghkhanpour@gmail.com
[2]Faculty of Electrical, Computer and IT Engineering,
Qazvin Azad University, Qazvin, Iran
O_sojoodi@qiau.ac.ir

ABSTRACT

Nowadays the determination of the dynamics of sequential data, such as marketing, finance, social sciences or web research has receives much attention from researchers and scholars. Clustering of such data by nature is always a more challenging task. This paper investigates the applications of different Markov models in web mining and improves a developed method for clustering web users, using hidden Markov models. In the first step, the categorical sequences are transformed into a probabilistic space by hidden Markov model. Then, in the second step, hierarchical clustering, the performance of clustering process is evaluated with various distances criteria. Furthermore this paper shows implementation of the proposed improvements with symmetric distance measure as Total-Variance and Mahalanobis compared with the previous use of the proposed method (such as Kullback–Leibler) on the well-known Microsoft dataset with website user search patterns is more clearly result in separate clusters.

KEYWORDS

Hidden Markov Model, distance metric, agglomerative clustering, categorical time series sequence, probability model

1. INTRODUCTION

Determining the dynamics in a Sequential data has become a critical step in many research fields. Current researches on data mining methods for dealing with big data in clustering of sequential data has recently aroused great interest[1] For example, discovering patterns in web navigation, similar to web mining, has become an important subject[3]. In this regard, [2] it was illustrated that traditional data mining approaches might be unsuitable for pattern discovery of websites users. Therefore, a large number of algorithms have been proposed for clustering web usage patterns. For example, the approaches that use K-means algorithm with KL distance metric as an alternative of Euclidian-metric distance [4], are resulted in the development of hierarchical model

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 125–136, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60211

based algorithms for web transactions clustering [5] and model-based approaches based on Markov models [6].

Along with data mining techniques, the use of probabilistic models such as Markov chain model [9], is useful in classification of web pages and generation of similarity and relation between different web sites. In order to conduct web mining, information from various sources such as web server access log, proxy server log, log browser, user profiles, data registration and meeting user transactions, cookies, bookmarks data , mouse clicks, surveys and other data can be collected as a result of an interaction.

In [10] it was investigated that the Markov models in web mining can be used to predict the user's next action; for example, using Markov models, social networks can predict future visits of users. Social networks can be mapped as a Markov chain; also using hidden Markov models with support vector machine classification methods, predicting sports, weather and social activities on Twitter was possible.

HMM is a machine learning algorithm used for pattern recognition in various applications (e.g. speech recognition, text and movement). The algorithm consists of two random processes. Hidden processes are not visible directly but indirectly can be deduced throughout the random process that produces a sequence of observations. Statistical methods such as Markov models can be employed to explore the behavior of transient (temporary) web data.

In section 2 we review previous researches on the application of different types of Markov models in various fields of web mining. Section 3 introduces the issue and the constraints involved in solving them, using previous methods. Section 4 describes the process of modeling and hierarchical clustering problem using different distance-metric criteria. Section 5 describes the standardized data set that contains records of web users' browsing history by introducing and applying the proposed method. Section 6 compares the quality of clustering, using different distance measures, and reports findings and results. Section 7 and 8 present future work and references list respectively.

2. REVIEW OF RESEARCH ON THE APPLICATION OF MARKOV MODELS IN WEB MINING

Traditional hierarchical clustering algorithms commonly used in clustering are somewhat impractical because it requires more storage and computation when the number of observations is large. K-means algorithm is considered as one of the most widely used algorithms in web mining. For the clustering of web users and user sessions, modeling studies based on Boolean (met / not met) or based on the frequency (number of each page visits) were adapted in web application. In other studies exploring the general sequence, sequence pattern mining techniques are used in order to reduce the computational complexity and produce significant clusters (meaningful). Using statistical models such as Markov models, in particular in the clustering process and display data encoding, is a more efficient way, so that the review of the current paper, Markov chain models role is well-appreciated with capabilities in three areas of web mining (usage, content, and structure mining). C. Xu et al., [13] proposed a hidden semi Markov model in web usage mining that the page sequence {page1, page2, … } can be described as a Markov chain; the HTTP requests $\{r_1, r_2, \dots, r_n\}$ or interval time $\{O_1, O_2, \dots, O_{n-1}\}$ between adjacent requests can be treated as the observations of Markov chain; and each Markov state can output multiple

observations continuously. Obviously, the user click behaviour conforms to hidden semi-Markov model (HsMM). Their methods were used state selection algorithm based on K-means clustering on backbone of a state China Telecom data set. Luca De Angelis et al., [14] proposed an extended hidden Markov model and time series in web usage mining. They mine categorical sequences from data using a hybrid clustering method that observation was sequences (variable length) of change transition between states and hidden states generated dynamics by time series, training algorithm was EM on well-known Microsoft dataset with website users search patterns. Sungjune Park et al., [15] proposed Markov chain in web usage mining. Parameters were page categories as state that each Markov chain represents the behaviour of a specific subgroup and categorize page with K-means & Kmeans + Fuzy ART methods, training algorithm was EM on Information Server (IIS) logs for **msnbc.com** data set for the entire day of September, 28, 1999. Yu-Shiang Hung et al., [16] proposed combined Markov model with ART2-enhance in web usage mining, each Markov chain represents the behaviour of a specific subgroup and training algorithm was EM on all of march, 2012, 3391 sessions of 157 elders data set were identified for analysis. Yi Xie et al., [17] proposed a Large-scale hidden semi Markov model for web security in web usage mining. Parameters were pages of web site as hidden state that states transition was the structure of web page links and inner requests of pages as observation. Training algorithm was unsupervised extended re-estimation [21,22] and entropy clustering on anomaly detection in behaviours of user navigation data set.

3. PROBLEM DEFINITION

As seen in the example of table 1, sequences may be restrictions on the issue of calculating the distance (similarity / dissimilarity) between them exist.

Table 1. Observed sequences A and B.

Time	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Sequence A	1	1	1	2	2	2	2	2	1	1	1	2	1	2	1	2
Sequence B	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2

Table 2. observed transitions between states.

Sequence A		
States	1	2
1	4	4
2	3	4

Sequence B		
States	1	2
1	4	4
2	3	4

Table 2 contains two categorical sequences of A & B with state- space {1 and 2} and length of 16. According to the table 1, the sequences are different. In fact, from a Markov Chain perspective these two sequences are identical as its sufficient statistics are the same.As it is shown in table 2, the starting condition and state of the state transition matrix are identical in observations. Therefore, based on these data, applying any distance-metric between A and B would be null; in other words, the sequences would be identified as identical which always results in the belonging to the same cluster.

The aim of this paper is to improve method that based on hidden Markov model in [14] clustering the search pattern of web users. The proposed improvements on method of this paper are a

combination of model-based clustering approach, which is built through the development of one HMM. In particular, the process of clustering with correlated data is observed, developed in time series and categorical data set is transformed to a probabilistic space in a way that the symmetric distance-metrics as Kullback–Leibler, Total-Variance and Mahalanobis could be applied on it.

The BCD algorithm clusters the time series based on the distances between the matrixes of observed transitions between states; while the procedure of clustering in this paper rests on the distances between posterior probabilistic of hidden states resulted from HMM after learning phase (estimated using Baum Welch algorithm[24]). Additionally, the estimation from HMM panel the similar hidden part of each of the time series lets the last probability of each sequence be comparable.

4. PROBLEM MODELING AND CLUSTERING

This section introduces a flexible method for categorical clustering of times series and clustering procedures based on model and hierarchical.

4.1. Definintion, Concepts and Notation

Y is a sample of n objects from time series sequences and each of its, with variable length (from 1 to T_i) that t={1, 2 , … , T_i}, denoted by Y = ($Y_1, Y_2, \dots, Y_n$) subject to $Y_i = (Y_{i\,1}, Y_{i\,2}, \dots, Y_{i\,T_i})$ and Z = ($Z_1, Z_2, \dots, Z_n$) explains different hidden states of Markov. φ Explains the set of parameters which is f(y_i; ϕ) the probability density function for object i with the particular parameter φ. The logarithm of maximum likelihood (ML) function of data for the set of parameters is $l(\varphi; y) = \sum_{i=1}^{n} \log f(y_{i;}\, \varphi)$.

$$f(Y, Z \mid \phi) = ? \tag{1}$$

$$\text{data set is } Y = \{ Y_1, Y_2, \dots, Y_n \} \tag{2}$$

$$Z = \{ Z_1, Z_2, \dots, Z_n \} \tag{3}$$

$$| Y_i | = T_i \;, Y_i = \{ Y_{i\,1} = A, Y_{i\,2} = B, \dots, Y_{i\,T_i} = \cdots \} \;, Y_{i\,j} \in \{ 1, \dots, M \} \tag{4}$$

$$| Z_i | = T_i \;, \; Z_i = \{ Z_{i\,1}, Z_{i\,2}, \dots, Z_{i\,T_i} \} \;, Z_{i\,j} \in \{ 1, \dots, K \} \tag{5}$$

4.2. Step 1 : HMM panel

The first step, provides a model-based approach through the development of the concept of hidden Markov model. It assumes that time observations (time series) Y_t is dependent on the hidden random process of Z_t which is defined with K states. The relation between data series is observed and the hidden process for object i is shown in picture 2.

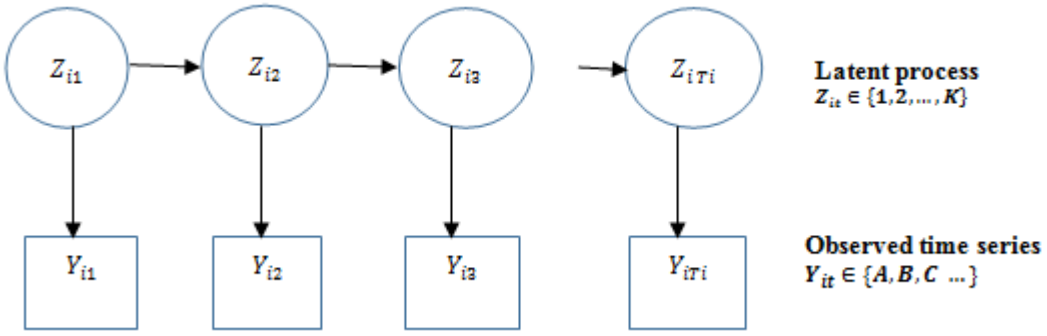

Fig. 2. Graphical representation of the model

The above graph shows the serial dependency in the observations which are completely affected by Z_t process; thus, HMM with the defined variable from different states of K for each time observation, HMM estimates the hidden variable for the entire T_i. For the set of ϕ parameters, HMM panel is determined for object i as follow:

$$\text{HMM } f(Y_i, Z_i|\phi) = f(y_{i1} | z_{i1}).\ f(z_{i1}) \prod_{t=2}^{T_i} f(y_{it} | z_{it}) f(z_{it} | z_{i(t-1)}) \quad (6)$$

$$\text{and } f(Y_i | \phi) = \sum_{z_i} f(Y_i, Z_i|\phi) \quad (7)$$

$$f(Y, Z | \phi) = \prod_{i=1}^{n} f(y_i, z_i | \phi) \quad (8)$$

$$\text{all i} : \lambda_i \equiv f(z_i = k) \ ; \ \sum_{i=1}^{n} \lambda_k = 1 \quad (9)$$

$$A \text{ (latent state transition matrix)} \equiv \pi_{wk} \equiv f(z_{it} = k \mid z_{it-1} = w); \sum_{k=1}^{K} \pi_{wk} = 1 \quad (10)$$

$$B \text{ (observe state transition matrix)} \equiv B_{kj} \equiv f(Y_{it} = j \mid z_{it} = k) \ ; \sum_{j=1}^{M} B_{kj} = 1 \quad (11)$$

Now, due to high number of parameters estimation of $\hat{\varphi} = argmax_{\varphi} l(Y | \phi)$ using ML method would be very difficult. Therefore, approximation method of EM, which is Baum Welch for HMM, will be used.

After finishing the learning phase of HMM, the set of $\hat{\varphi}$ ($\widehat{\lambda_k}$ ،$\widehat{\pi_{wk}}$ و $\widehat{B_{kj}}$) parameters will be estimated and they let us calculate the posterior probabilities. The $\widehat{u_{ik}}(t) = f(\ i \in k \text{ at time } t \ | y_i, \hat{\varphi})$ which is the probabilities of an observation and is estimated in a hidden states in time *t* condition on time series observations and estimated parameters. In that paper, transformation of the main dataset to $\widehat{u_{ik}}(t)$ posterior probabilities is suggested. This transformation provides two advantages: 1. Posterior probabilities contain exclusive information of each time series and this could be used simply in clustering. 2. They can be compared; for instance, for $k = 1, 2, \dots, K$ and $\widehat{u_{ik}}(t)=1$, $i = 1, 2, \dots, n$. The most important analytical step of this paper is the determination of hidden conditions (K) of Markov process as in [14]. The value of K is identified 12 in the implementation of web mining on msn web data set(available on kdd.ics.uci.edu).

4.3. Step 2: Hirachical Agglomerative Clustering

The second step is to extract the information provided by the probabilities obtained in the first step to determine clusters of sequences characterized by similar dynamic patterns. Specifically, if two time series with posterior probability are assigned (allocated) to the same hidden status, then they should be in the same cluster. Natural chance to calculate the distance between two probability distributions will be KL.

All t,i : $f(z_{it(1..T_i)} = k \mid Y_{i(1..T_i)})$ (12)

Table 3. Mapping i'th sequence to form the posterior probabilities

	1	2	…	t	…	
1						
2				…		
K			…		…	
...				…		
K						

4.4. Distance Metric Improvement in Clustering Phase

Instead of KL similarity measure, other criteria those are applicable to the data with statistical distribution, such as Total Variance and Mahalanobis, can be used.

4.4.1. TV Distance Metric [11]

$q = \widehat{u_{jk}}$, $p = \widehat{u_{ik}}$ (and the opposite) $D_{TV}(p \parallel q) = ½ \sum_{i=1}^{M} | p(i) - q(i)|$ (13)

4.4.2 Mahalanobis Distance Metric [12]

This distance criteria for an observation $Y = (y_1, y_2, \dots, y_N)^T$ from a group of observations with the average observation $\mu = (\mu_1, \mu_2, \dots, \mu_N)^T$ and covariance matrix S is defined as follows:

$$D_M(Y_i, Y_j) = \sqrt{(Y_i - \mu)^T S^{-1} (Y_j - \mu)} \quad Y = (y_1, y_2, \dots, y_N)^T \tag{14}$$

The steps of implementation of Mahalanobis distance measure by using principal component analysis PCA [11]:

- Calculation of the total length of time series sequence data: $T_{total} = \sum_{i=1}^{n} T_i$

- With concatenation of probability matrix for each of the sequences, the total matrix has the dimension of the $T_{total} * K$ is formed (K number of hidden states).

- execution of PCA feature extraction algorithm on the previous step matrix (Without dimension reduction)

  - eigenvectors matrix (each column is a special-vector in descending order)

  - the transformed total matrix of the PCA space

  - variance in the directions of the eigenvectors

- To be in line with the amount of variance should be divided every direction, the standard deviation of direction.

5. WEB USAGE MINING

In this section, using the proposed method, msnbc.com (available in kdd.ics.uci.edu) dataset have been analysed with many distance measurements. The application aims to analyze the sequence of pages requested by the user search patterns visits to determine the clusters identified through different websites and search behavior on the Web. This data set is also used by other researchers [18, 19].

5.1. Data Set Description

Dataset includes a record number of 989,818 (each record a sequence of different pages a user visits) which is recorded by Microsoft MSN for websites visitors during a full day (28 September 1999). The variety of the visited web pages are 1 to 17:

(1) frontpage, (2) news, (3) tech, (4) local, (5) opinion, (6) on-air, (7) misc, (8) weather, (9) health, (10) living, (11) business, (12) sports, (13) summary, (14) bbs (bulletin board service), (15) travel, (16) msn-news, and (17) msn-sports.

Z considered hidden states of model of HMM, Y is Set of sequences of variable length, each of the constituent elements of each sequence in the moment t (t is 1 to T_i) have one of the 17 categories may be. The analysis and application of the proposed method in a random sample of about one percent of the data that has at least two pages have (have at least one transition) sample size n = 6244 are considered.

5.2. Step 1: HMM panel

In the first step of HMM panel, for reaching the optimum number of hidden state, K with values of 1 to 15 is analyzed and in order to avoid local maxima, the algorithm is run for 100 times. The best K (the number of hidden states for each item in the sequence) was determined 12.

5.3. Step 2: Hierarchical Agglomerative Clustering with Different Distance Metrics

Now we compute the distance between the two sequences with different distance metrics , then in the step 2, we start to cluster the obtained distance matrix with the usage of hierarchical agglomerative clustering algorithm with different distance method between clusters such as complete linkage method. As can be seen in the following figures (3, 4, 5, 6 and 7) according to the dendrograms' using of the different distance metrics, we are cutting the tree with the biggest cut, one of cluster that consists a large number of users, include the users who prefer short meetings with specific topics such as news headlines a few clicks, weather and sports search. Thus we could label this cluster as specialists' users. The second cluster represent generalist users such as web users have longer sessions characterized by longer sequences of clicks and prefer different topics on the website.

6. EVALUATION AND RESULTS

Applying KL, TV and Mahalanobis distance criteria with different distance methods from the figures of part 5.3 it is concluded that the implementation of the second part of the algorithm with the criteria of TV-ward, TV-complete, Mahalanobis-complete (with respect to the maximum distance between levels of clustering) respectively have more clearly result in separate clusters than distance measure (KL) utilized in the referenced paper in the clustering level.

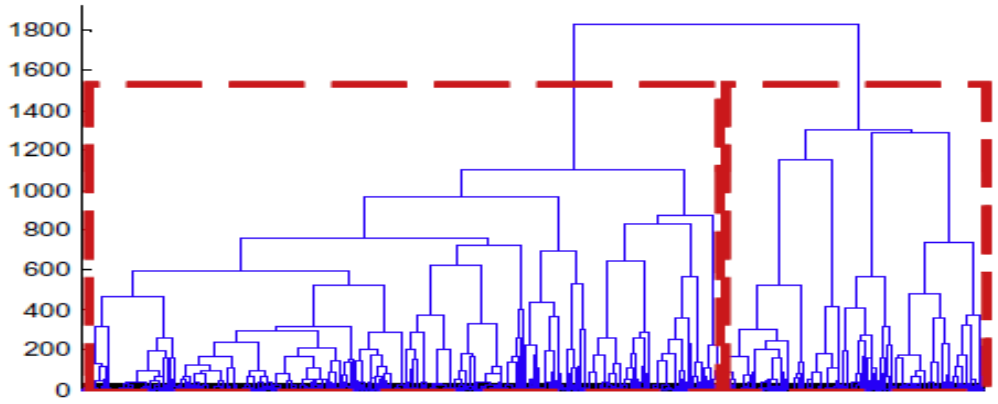

Fig. 3. Dendrogram -using KL with complete method (Basis for comparison- *fourth choice*)

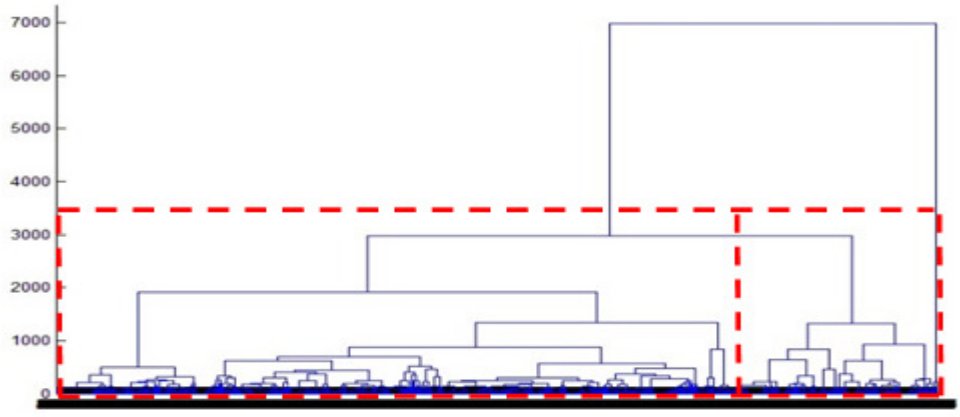

Fig. 4. Dendrogram -using TV with ward method (The best of the three criteria-*first choice*)

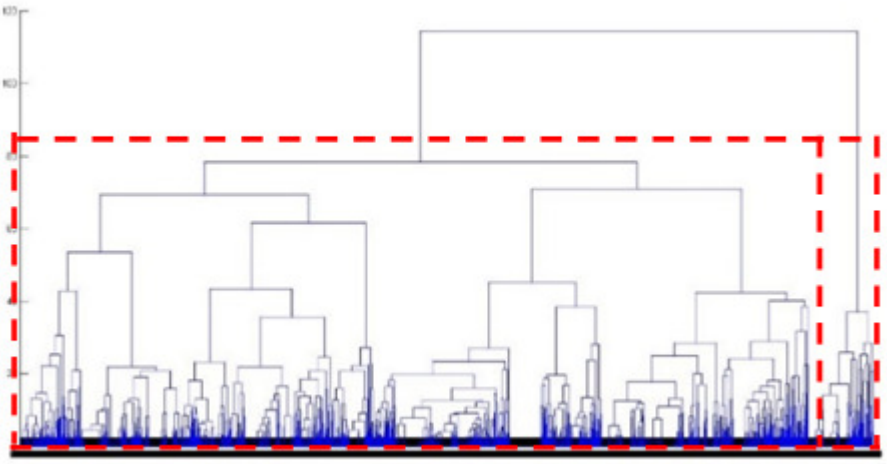

Fig. 5. Dendrogram -using TV with complete method(*second choice*)

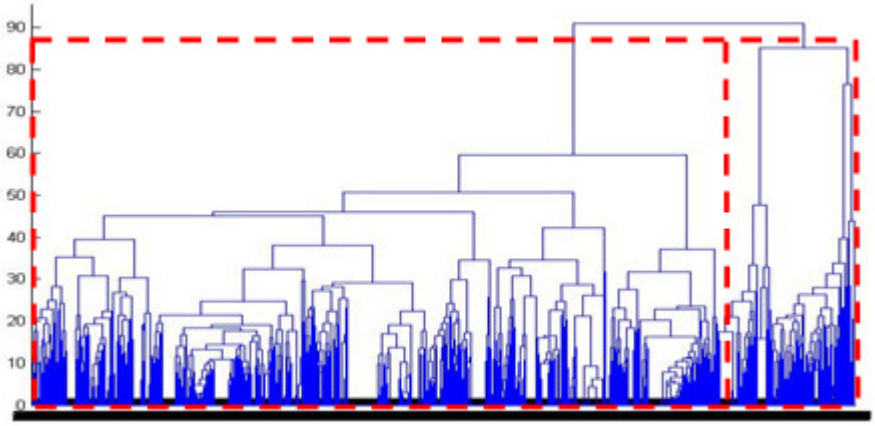

Fig. 6. Dendrogram -using Mahalanobis with complete method(better than KL – *third choice*)

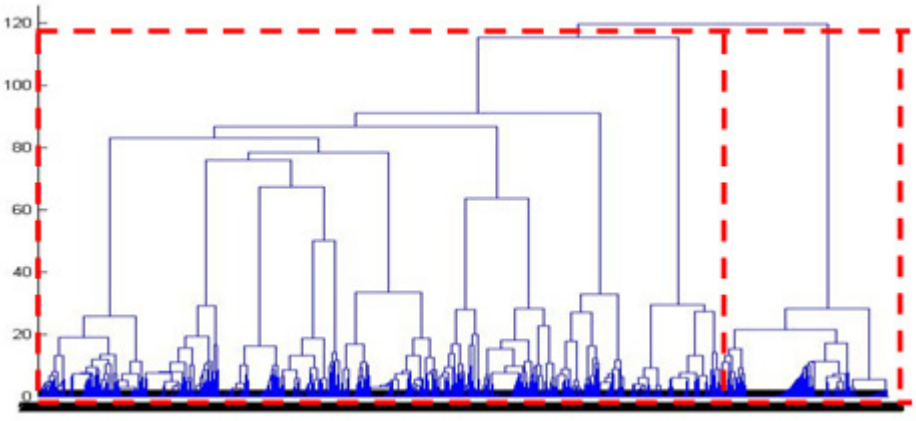

Fig. 7. Dendrogram -using Mahalanobis with complete method (*worse than KL and thired choice*)

7. CONCLUSIONS

In this paper we have improved a method based on hidden Markov model for clustering web users with different symmetric distance measures. Among the measures discussed, as Total-Variance and Mahalanobis compared with the previous work that was used of the proposed method (such as Kullback–Leibler) are better results and more clearly separated clusters as test results show up.

8. FUTURE WORKS

In order to development of the current work, we can do the next phase HMM training and clustering, we can predict the next sequence that improvement attractive website for website designers, and predict future patterns of search engine users may be fruitful. The clustering is done using a combination of improved methodology presented in this paper for other applications, such as browsing behaviour anomaly detection website user [17] and also on the results of other data collection including the development of the proposed method in this paper.

REFERENCES

[1] Ananthanarayana, V. S., Murty, M. N., & Subramanian, D. K. (2001). Efficient clustering of large data set. Pattern Recognition, 34, 2561–2563.

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

[3] Vakali, A., Pokorny, J., & Dalamagas, T. (2004). An overview of Web data clustering practices (pp. 597–606). Berlin: Springer.

[4] Petridou, S. G., Koutsonikola, V. A., Vakali, A. I., & Papadimitriou, G. I. (2006). A divergence oriented approach for Web users clustering. In M. e. a. Gavrilova (Ed.), ICCSA 2006. LNCS 3981 (pp. 1229–1238). Heidelberg: Springer.

[5] Yang, Y., & Padmanabhan, B. (2011). Generalized Markov models of infectious disease spread: A novel framework for developing dynamic health policies. European Journal of Operational Research, 215(3), 679–687.

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

[7] Ramoni, M., Sebastiani, P., & Cohen, P. (2002). Bayesian clustering by dynamics. Machine Learning, 47(1), 91–121.

[8] Ramoni, M., Sebastiani, P., & Kohane, I. (2002). Cluster analysis of gene expression dynamics. Proceedings of the National Academy of Sciences of the United States of America, 99, 9121–9126.

[9] V.Chitraa, Dr. Antony Selvdoss Davamani, A Survey on Preprocessing Methods for Web Usage Data Information Retrieval System, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010

[10] Dias, J. G., Cortinhal, M. J. (2008). The skm algorithm: A k-means algorithm for clustering sequential data. In Geffner, H., Prada, R., Alexandre, I. M., David, N.(Eds.), Proceedings of the advances in artificial intelligence – Iberamia. Lecture notes in computer science (vol. 5290, pp. 173–182).

[11] Sergios Theodoridis, Pattern Recognition, 4th edition, Copyright © 2009, Elsevier Inc. All rights reserved.

[12] Bishop, Pattern Recognition and machine learning, 2end edition, 2006 Springer Science, Business Media, LLC

[13] C. Xuchuan, C. Dua, G.F. Zhao, S. Yu, A novel model for user clicks identification based on hidden semi-Markov, Journal of Network and Computer Applications 36 (2013) 791–798, ELSEVIER.

[14] Luca De Angelis, José G. Dias,Mining categorical sequences from data using a hybrid clustering method, European Journal of Operational Research xxx (2013), ELSEVIER.

[15] Sungjune Park, Nallan C. Suresh, Bong-Keun Jeong, Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm, Data & Knowledge Engineering 65 (2008) 512–543, ELSEVIER.

[16] Yu-Shiang Hung, Kuei-Ling B. Chen, Chi-Ta Yang, Guang-Feng Deng, Web usage mining for analysing elder self-care behavior patterns, Expert Systems with Applications 40 (2013) 775–783, ELSEVIER.

[17] Yi Xieyicn and Shun-Zheng , A Large-Scale Hidden Semi-Markov Model for Anomaly Detection on User Browsing Behaviors, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 17, NO. 1, FEBRUARY 2009.

[18] Dias, J. G., & Vermunt, J. K. (2007). Latent class modeling of website users' search patterns: Implications for online market segmentation. Journal of Retailing and Consumer Services, 14(6), 359–368.

[19] Ramos, S., Vermunt, J., & Dias, J. (2011). When markets fall down: Are emerging markets all the same? International Journal of Finance and Economics, 16(4), 324–338.

[20] Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6(2), 461–464.

[21] M. Kantardzic, WEB MINING A ROADMAP, Methods And Algorithm. New York: IEEE Press, 2009.

[22] X.Yi and Y. Shunzheng, "A dynamic anomaly detection model for web user behavior based on HsMM," in Proc. 10th Int. Conf. Computer Supported Cooperative Work in Design (CSCWD 2006), Nanjing, China, May 2006, vol. 2, pp. 811 816

[23] Bamshad Mobasher, 'Data Mining for Web Personalization' , School of Computer Science, Telecommunication, and Information Systems DePaul University, Chicago, Illinois, USA

[24] Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. IEEE Information Theory Society Newsletter, 53:1(4), 10–13.

AUTHORS

First Author Sadegh Khanpour is PhD student of Qazvin Islamic Azad University, as well as the director of the data mining research and development team from one of the chain stores in Iran.

Second Author Omid Sojoodi holds a PhD in Artificial Intelligence from the UPM University of Malaysia, and faculty member of Islamic Azad University of Qazvin.