A Probabilistic Validation Algorithm for Web Users' Clusters [*]

George Pallis, Lefteris Angelis, Athena Vakali
Department of Informatics
Aristotle University of Thessaloniki,
54124, Thessaloniki, Greece
gpallis@ccf.auth.gr, {lef,avakali}@csd.auth.gr

Jaroslav Pokorny
Faculty of Mathematics and Physics
Charles University
Praha, Czech Republic
pokorny@ksi.mff.cuni.cz

Abstract – *Cluster analysis is one of the most important aspects in the data mining process for discovering groups and identifying interesting distributions or patterns over the considered data sets. In the context of Web data mining, model-based clustering algorithms are often used to cluster similar users' sessions in order to determine Website access behaviors. An important issue in cluster analysis is the evaluation of clustering results to find the partitioning that best fits the underlying data. In this paper, we present a novel validation technique for model-based clustering approaches.*

Keywords: Web Data Clustering, Cluster Validity, Data Mining and Management.

1 Introduction

As the number of Web users and the number of accessible Web pages grows significantly, it is becoming increasingly difficult for users to find documents that are relevant to their particular needs. Users must either browse through a large hierarchy of concepts to find information or submit a query to a publicly available search engine (spending a lot of time through hundreds of results, most of them irrelevant). Therefore, the process of understanding the users' navigation behavior is challenging but fundamental in improving Web query answering, link structure and in simplifying navigation through a large number of individual Web pages.

In this context, clustering is expected to provide a general grasp of the Web for effective Web users' navigation and searching. In general, clustering is one of the most important practices in the data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data [13]. The clustering problem is about partitioning a given data set into clusters (groups) such that the data points in the cluster are more similar to each other than points in different clusters. In the context of Web mining, there are two kinds of clusters with special interest: users' sessions clusters and page clusters. Clustering of users' sessions identifies groups of users exhibiting similar browsing patterns. More specifically, a session provides information about the sequence of pages viewed by a user as (s)he moves through a Web site and each one reflects an individual user's behavior. Such knowledge is especially useful for customizing a Web site to the needs of a particular user or a set of users [15]. On the other hand, clustering of Web pages tends to establish groups of pages based either on their content or on their hyperlink information [14], [16].

In this paper, we focus on validating the Web users' clusters. Specifically, several clustering approaches have been proposed in the past, assigning the sessions (users' behaviors) with common characteristics into the same cluster [16], [17], [18]. These may be classified into two schemes:

- *Similarity-based*: It uses distance functions (e.g. Euclidean, Manhattan, cosine etc.) to measure similarities among sessions [2], [20]. Distance functions can be determined either directly, or indirectly, although the latter is more common in most applications. Hierarchical and partitional are the most indicative approaches that belong to this category. More specifically, hierarchical schemes use a distance function to determine a hierarchy of clustering, merging always the most similar clusters. On the other hand, the partitional algorithms determine a "flat" clustering into a specific number of clusters (e.g. K-means, K-mode etc.) [4].

- *Model-based*: Each cluster is represented by a probability model and the sessions are partitioned according to the order in which users request Web pages. More specifically, each cluster has a data-generating model with different parameters for each cluster. Model-based schemes are usually preferred from the Web community since they can efficiently describe the dynamic evolution of the Web [1], [21]. In this paper, we also use a model-based algorithm to cluster the users' sessions.

[*] **0-7803-8566-7/04/$20.00 © 2004 IEEE.**

The rest of the paper is organized as follows. In Section 2, we provide a brief discussion of related work, and clarify the contributions of this paper. In Section 3, we describe the Web Logs and the users' session identification process. In section 4, we present the clustering algorithm that is used to group the users' sessions. In Section 5, we present a novel validation tecnhique for model-based clustering schemes. In Section 6, experimental results are given. Finally, we conclude the paper and give some future remarks.

2 Related Work and Paper's Contribution

The problem of evaluating the results of a clustering algorithm is one of the most important issues in cluster analysis and has attracted research interest [9], [11], [22]. The problem is related to the question: After applying a cluster algorithm, how can one assess the quality of the clusters returned? Clustering schemes always produce a partition of the given data set although there may be no real clusters on the data distribution. In order to select the validation procedure, the system searches for the optimal parameters' values for a specific clustering algorithm so as to result in a clustering scheme that best fits its data. Many of the most popular clustering algorithms require their parameters to be tweaked for the best results, but this is impossible if one cannot assess the quality of the output. Cluster validity has been proposed in the literature [11] based on several criteria. More specifically, these can be categorized into three approaches, as follows:

- *External approach*: It evaluates the results of a clustering method based on a pre-specified structured on a data set, which reflects a user's intuition about the clustering structure of this data set.

- *Internal approach*: It evaluates the clustering result in terms of quantities obtained from the data set itself. This approach is used by the authors in [12] in order to validate clusters of users' sessions. In particular, their method provides only a sense about how similar are the sessions within the cluster. This is indicated by using the Frobenius norm of the differences between the sessions and the cluster's mode. Our work also falls in this approach.

- *Relative approach*: It compares the evaluation of a clustering structure by other clustering schemes, modifying only the parameter values.

The main technical contribution of the paper can be summarized in the following:

- Suggestion of a validation algorithm for model-based clustering. This algorithm is based on a statistical chi-square test (χ^2) [19] employed to each cluster.

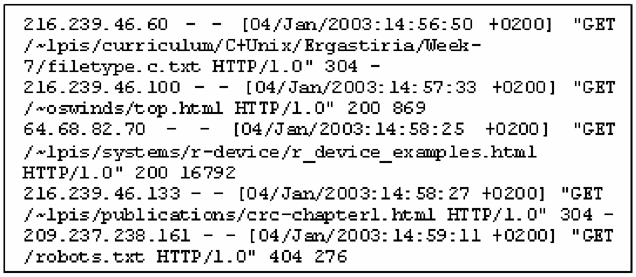
- The statistic criterion provided does not depend on tunable parameters.

- The algorithm was tested on a real data set collected from an educational Web server (the Department of Computer Science in Aristotle University of Thessaloniki).

To the best of our knowledge this is the first approach dealing with validation for model-based clustering schemes. Specifically, these schemes have been widely used for describing the dynamic evolution of the Web and have shown promising results in many Web applications. Therefore, a validation algorithm for model-based schemes may offer new perspectives for an efficient model-based clustering approach.

3 Web Logs and Users' Access Sessions

Each user that visits a Web server leaves a "trace" in the server log file. The trace consists of logging some user information (client IP address), date/time of request, individual page requested, success of the operation, etc. The useful information about the traffic on the server is stored in a large server log file. Popular Web servers' log files get millions of lines every day. Figure 1 presents a sample of a Web server log file. These data are undergone a certain pre-processing, such as invalid data cleaning and session identification [6]. Data cleaning removes the records which do not include useful information for the users' behavior, such as graphics, javascripts, small pictures of buttons, advertisements etc.



```
216.239.46.60  -  -  [04/Jan/2003:14:56:50 +0200]  "GET
/~lpis/curriculum/C+Unix/Ergastiria/Week-
7/filetype.c.txt HTTP/1.0" 304 -
216.239.46.100 - - [04/Jan/2003:14:57:33 +0200] "GET
/~oswinds/top.html HTTP/1.0" 200 869
64.68.82.70  -  -  [04/Jan/2003:14:58:25  +0200]  "GET
/~lpis/systems/r-device/r_device_examples.html
HTTP/1.0" 200 16792
216.239.46.133 - - [04/Jan/2003:14:58:27 +0200] "GET
/~lpis/publications/crc-chapter1.html HTTP/1.0" 304 -
209.237.238.161 - - [04/Jan/2003:14:59:11 +0200] "GET
/robots.txt HTTP/1.0" 404 276
```

Figure 1. A sample of Web server log file

In this paper, the remaining page requests are categorized into V different categories. The process of grouping the Web pages into categories is a usual practice, since it improves the data management and in addition eliminates the complexity of the underlying problem (since the number of page categories is smaller than the number of Web pages in a Web site) [1], [3], [12]. In particular, the individual pages are grouped into semantically similar groups. Scanning for specific keywords that occur in the URL string of page request makes the assignment of the page requests to a category.

In order to identify the users' sessions, heuristic methods are usually used based on IP and session time-outs [5]. In this paper, we first consider that we have an ordered set of traces with respect to the IPs. Therefore, a new session is created when a new IP address is encountered or if the visiting page time does not exceed 30 minutes for the same IP address.

4 Clustering Users' Sessions

The adopted model-based assumes that the data (based on the log files) are generated as follows:

- A user arrives at the Web site in a particular time and is assigned to one of the underlying clusters with some probability. The number of clusters may be determined by using several probabilistic models, such as BIC (Bayesian Information Criterion), bayesian approximations, or bootstrap methods [10].

- The behavior of each cluster is governed by a statistical model and the user's behavior is generated from this model to that cluster.

In general, each cluster has a data-generating model with different parameters for each one. Therefore, this model can be well defined, if only we learn all the parameters of each model component: the probability distribution used to assign users to the various clusters and the number of components. Once the model is learned, we can use it to assign each user to a cluster or fractionally to a set of clusters. The parameters can be learned using the Expectation-Maximization (EM) algorithm. The EM algorithm originates from [8] and in [3] a method for employing on EM on users' sessions is proposed. In particular, the EM algorithm is an iterative procedure that finds the maximum likelihood estimates of the parameter vector by repeating the following steps:

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

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

The two steps are iterated until the iterations converge.

In this paper, we also cluster users by learning a mixture of first-order Markov models using the EM algorithm. Specifically, Markov models can be viewed as stochastic generalizations of finite-state automata, when both transitions between states and generation of output symbols are governed by probability distributions. In our framework, we consider a Markov chain model where we model probability that user will go to a certain page category given (s)he is viewing the current page category. Therefore, we have a transition matrix of size $V \times V$ (where V is the number of categories) and a set of V initial probabilities describing how likely is that user will begin his/her navigation session in a given page category. To model heterogeneity of users we use a mixture of first-order Markov chains, where each component in a mixture represents a behavior described by a single Markov chain. Concerning the complexity of the EM algorithm, it depends on the complexity of the E and M steps at each iteration [3]. For example, in our case (Markov mixtures) the complexity is linear in the sum of the lengths of all sessions. Note, that for more complex mixture models the complexity can be higher. Once the model is specified, we use the EM algorithm and probabilistic out-of-sample evaluation to determine the best number of components. A model is fitted on a subsample of sessions (the so-called training data set) and then scored on the remaining data (the so-called testing data set). Thus, we get an objective measure of how well each model fits the data. The model with the minimum out-of-sample predictive log score is selected.

5 Clustering Validation Algorithm

The clustering algorithm described above results in a number of clusters, where each cluster is comprised from users' navigation sessions. We assume that each cluster can be represented by a first-order Markov ergodic chain. By the term "ergodic", we mean a Markov chain that has the following two properties:

- Each node can reach any other node (all states intercommunicate),

- The chain is not periodic (all states have period one)

Each first-order Markov chain corresponds to an individual transition matrix (which contains the transition probabilities among the states) and a vector (which represents the initial state probabilities). In this context, it should be emphasized that the method following is valid only if all clusters are represented by ergodic Markov chains. However, a large number of experiments conducted by the EM algorithm, it has been observed that the above condition is always satisfied [1], [3].

In order to validate the clustering scheme, we consider the equilibrium distribution of each cluster produced by the algorithm. These distributions represent the probabilities of a user to access each state in infinite number of states independently of its initial state. We believe, that the equilibrium distribution offers a complete and objective view for the navigation behavior of Web users.

Theorem 1: If P is the transition matrix of a homogeneous ergodic Markov chain, then there is a unique vector f=(f_1, …, f_V), such that

$$\lim_{n\to\infty} P^n = \begin{pmatrix} f \\ f \\ \dots \\ f \end{pmatrix} \qquad (1)$$

A thorough study and classification of finite Markov chains and the proof of this theorem is given in [7]. This theorem offers us a way of approximately evaluating the access frequencies of the nodes, by simply calculating powers of the transition matrix. It gives us a way to evaluate the relative frequency of accessing (retrieving) nodes 1, ..., V respectively in a long run, based on the transition probabilities of the initial browsing graph. It is known that in the theory of stochastic processes the vector f is called the equilibrium or stationary distribution of the Markov chain since any element represents the limiting probability of accessing the respective nodes 1, …, V after infinite number of steps.

Then, the validation is performed by testing the homogeneity of the equilibrium distribution by the χ^2 test.

Table 1. A Contingency Table for Chi-square Testing

	States				
Clusters	**A_1**	**A_2**	**…**	**A_V**	**sum**
C_1	O_{11}	O_{12}	…	O_{1V}	Y_1
C_2	O_{21}	O_{22}	…		Y_2
…	…	…	…	…	…
C_K	O_{K1}	O_{K2}	…	O_{KV}	Y_K
sum	X_1	X_2	…	X_V	S

More specifically, χ^2 testing [19] is used to test the homogeneity among multiple clusters with probabilistic distributions by constructing a contingency table. This statistic is used to assess evidence that two or more distributions are dissimilar. Considering that in model-based approach the clusters represent a probabilistic distribution, we can directly apply the test of homogeneity by fitting the state frequencies in the cluster into the contingency table, which reflects the fact that our modeling simplifies the testing. Formally, we assume that there are K clusters C_1, C_2, ..., C_K and each of them is generated from its own probability distribution. Moreover, there are V different states altogether denoted by A_1, A_2, ..., A_V. In our framework, the states represent the page categories. Table 1 is the contingency table for testing. A contingency table test (or test of independence) is one that tests the hypothesis that the data are cross-classified in independent ways. In particular, O_{ij} stands for the frequency of A_j state in cluster C_i. O_{ij} is computed by multiplying the relative frequency of A_j state with the number of sessions that belong to cluster C_i. X_i is the sum of all the O_{ij} in ith column and Y_j is the sum of all the O_{ij} in jth raw. In this framework, we want to test the following hypothesis (for all the states and clusters of the underlying model):

Null Hypothesis (H_o): The distributions of the states in each cluster are all the same.

Testing: The following equation computes the χ^2 statistic:

$$\chi^2(C_1, C_2, ..., C_K) = \sum_{i=1}^{K}\sum_{j=1}^{V}\frac{\left(O_{ij} - Y_i \times \frac{X_j}{S}\right)^2}{Y_i \times \frac{X_j}{S}} \qquad (2)$$

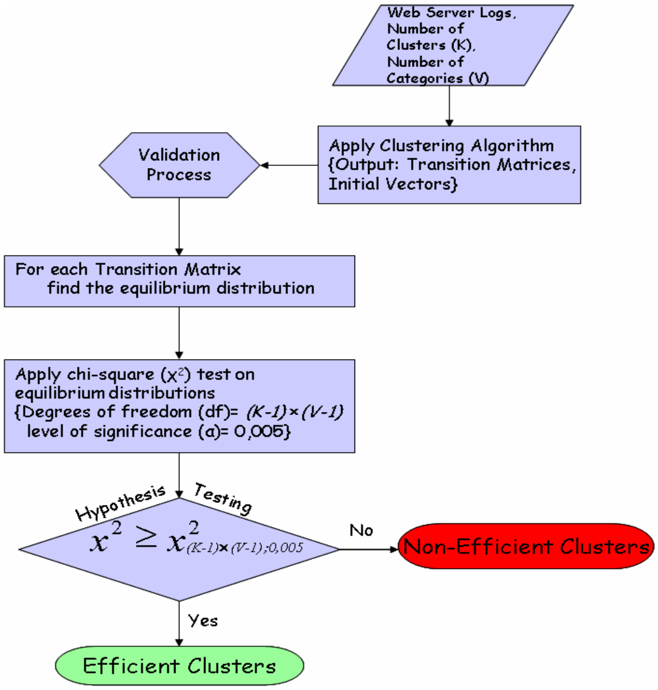

Web Server Logs,
Number of
Clusters (K),
Number of
Categories (V)
Apply Clustering Algorithm
{Output: Transition Matrices,
Initial Vectors}
Validation
Process
For each Transition Matrix
find the equilibrium distribution
Apply chi-square (χ²) test on
equilibrium distributions
{Degrees of freedom (df)= (K-1)×(V-1)
level of significance (α)= 0,005}
Hypothesis Testing
$x^2 \geq x^2_{(K-1)\times(V-1);0,005}$
No
Non-Efficient Clusters
Yes
Efficient Clusters

Figure 2. A Flow-Chart for the Clustering Validation Algorithm

A large value of the χ^2 criterion (equation 2) shows that the equilibrium distribution for each cluster are significantly different, which in turn is an indication of the heterogeneity among clusters. Therefore, we should know a critical χ^2 value that is the boundary of the area of hypothesis' s rejection in a contingency table test. In order to find this critical value, we should define the level of significance (α) and the degrees of freedom (df). In statistics, it is known that a χ^2 has asymptotically a χ^2 distribution with (K-1)×(V-1) df [19]. Therefore, if the value of χ^2 distribution is greater than a critical value, such as $\chi^2_{(K-1)\times(V-1);a}$, we reject the H_o at the α level of significance. Otherwise, we fail to reject H_o. Figure 2 shows the proposed clustering validation algorithm in a flow-chart.

6 Experimentations

Web usage data from www.csd.auth.gr (an educational Web server in Greece) are used to test and validate the performance of our validation method proposed.

In our experiments, the number of categories is 9 (such as research, faculty, information etc.) and each category includes a number of URLs. The data set consists of approximately 3,000 users' sessions, with an average of 3,3 page views per session. Then, we select some of the sessions as (80% of the total data) training data set and the rest as testing data set [3].

Initially, we should identify a good value for the number of clusters. As we referred in Section 4, we choose the number of clusters by finding the choice that minimizes the out-of-sample predictive score. Figure 3 shows several out-of-sample log-likelihoods for varying number of clusters. From this figure, it is evident that the lowest value is for the choice of 6 clusters.

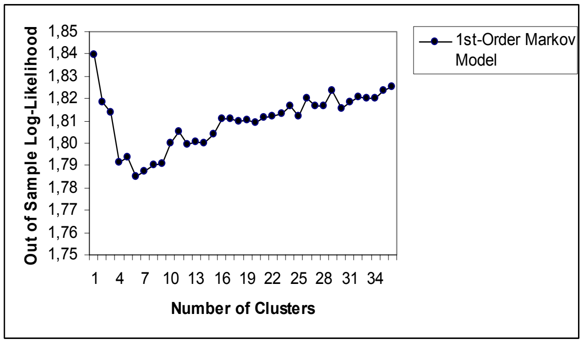


Figure 3. Number of Clusters

Then, we cluster the users' navigation sessions as described in Section 4. Each cluster is represented by a probabilistic distribution (first-order Markov model). In this context, we find the equilibrium distribution for each cluster, and then we apply the χ^2 test on them (as described on the previous Section). Table 2 presents the contingency table for our data set. Based on the specific data set, we result (using the equation 2) that the χ^2 = 2142,1 with 40 df. Considering that the $\chi^2_{40;0,005} = 66,76$, we conclude that the derived clusters have very different characteristics with each other. Thus, the resulted clusters are an effective choice.

Table 2. A Contingency Table Test

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	sum
C_1	27,57	17,78	102,10	136,94	120,12	13,09	7,81	9,86	26,74	462,00
C_2	105,21	9,97	46,14	321,32	230,57	9,83	9,99	9,31	44,67	787,00
C_3	18,81	11,96	10,28	19,04	18,57	9,04	4,52	8,92	53,87	155,00
C_4	9,11	6,52	6,84	14,57	16,26	6,40	8,43	6,41	13,45	88,00
C_5	41,65	32,61	32,88	34,92	996,29	31,36	34,61	32,47	45,21	1.282,00
C_6	20,28	17,88	18,51	19,99	23,11	17,62	106,50	17,23	23,89	265,00
sum	222,62	96,72	216,76	546,77	1.404,92	87,32	171,86	84,20	207,82	3.039,00

7 Conclusion and Future Work

In this paper, we have proposed a novel clustering validation algorithm for model-based clustering schemes, where each cluster is represented by a mixture of First-order Markov models. In general, First-order Markov models are reasonable choice for modeling users' navigation sessions. Mixture of Markov chains describes the data better showing that there are different prototypical behaviors of users. In this framework, EM algorithm can be applied to learning the Markov chains mixture model and it scales linearly with both number of mixture components and number of users' sessions.

Furthermore, the Markov models may provide valuable information for users' navigation behavior but it is often hidden. Statistical analysis helps to explore this hidden information in order to enhance the Web performance. For example, further analysis of the contingency table (such as correspondence analysis) or other measures of association can reveal interesting relations among clusters and states. For the future, we plan to investigate these issues and develop a novel algorithm for clustering Web users' sessions, based on some interesting characteristics of the Markov models.

References

[1] P. Baldi, P. Frasconi, and P. Smyth, "*Modeling the Internet and the Web*", Wiley Press,USA, 2003.

[2] A. Banerjee, and J. Ghosh, "Clickstream clustering using weighted longest common subsequences", Proc. of the International Workshop on Web Mining, SIAM Conference on Data Mining, pp.33 – 40, Chicago, USA, April 2001.

[3] I. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Model-based clustering and visualization of navigation patterns on a Web sSite", *Journal of Data*

Mining and Knowledge Discovery, Vol. 7, No. 4, pp. 399-424, 2003.

[4] S. Chakrabarti: "*Mining the Web*", Morgan Kaufmann Press, 2003.

[5] Z. Chen, A. Fu, and F. Tong, "Optimal algorithms for finding user access sessions from very large Web logs", *World Wide Web: Internet and Information Systems*, Vol. 6, pp. 259-279, 2003.

[6] R. Cooley, B. Mobasher, and J. Srivastava: "Data preparation for mining World Wide Web browsing patterns", *Knowledge Information Systems*, Vol. 1, pp. 5-32, 1999.

[7] D.R. Cox, and H. D. Miller, "The theory of stohastic processes", Chapman and Hall Press, 1997.

[8] A. P. Dempster, N. M. Lsird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Statistics Society B, Vol. 39, pp. 1-22, 1977.

[9] V. Estivill-Castro, "Why so many clustering algorithms - a position paper" *Journal of ACM SIGKDD Explorations*, Vol. 4, No. 1, pp. 65-75, 2002.

[10] C. Fraley and A. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis", *Computer Journal*, Vol. 41, pp. 578--588, 1998.

[11] M. Halkidi, Y. Batistakis, and M. Vazirigiannis, "On clustering validation techniques", *Journal of Intelligent Information Systems*, Vol. 17, pp. 107-145, 2001.

[12] Z. Huang, J. Ng, D. W. Cheung, M. K. Ng, and W. Ching, "A cube model for Web access sessions and cluster analysis", Proc. of the 3rd Internation Workshop on Mining Web Data (WEBKDD 2001), San Francisco, August 2001.

[13] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, Sep. 1999.

[14] T. Masada, A. Takasu, and J. Adachi, "Web page grouping based on parameterized connectivity" Proc. of the 9th International Conference on Database Systems for Advances Applications (DASFAA), Jeju Island, Korea, March 2004.

[15] B. Mobasher, R. Cooley, and J. Sristava, "Automatic personalization based on Web usage mining", *Communications of the ACM*, Vol. 43, No.8, pp.142-151, 2000.

[16] R. R. Sarukkai, "Link prediction and path analysis using Markov Chains.", *Journal of Computer Networks*, Vol. 33, pp. 377-386, 2000.

[17] R. Sen, and M. H. Hansen, "Predicting a Web user's next request based on log data", *Journal of Computations Graph Statistics*, Vol. 12, 2003.

[18] C. Shahabi, A. M. Zarkesh, J. Adibi, and V. Shah, "Knowledge discovery from users Web page navigation", Proc. of the 7th International Workshop on Research Issues in Data Engineering (RIDE'97), England, April 1997.

[19] G. Snedecor, and W. Cochran, "*Statistical methods*", Eighth Edition, Iowa State University Press, 1989.

[20] J. Xiao, and Y. Zhang, "Clustering of Web users using session-based similarity measures", Proc. of the International Conference on Computer Networks & Mobile Computing (ICCNMC'2001), Beijing, China, pp. 812-817, October 2001.

[21] A. Ypma, and T. Heskes "Categorization of Web pages and user clustering with mixtures of hidden Markov models", Proc. of the 4th International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles (WEBKDD 2002), pp. 31-43, Edmonton, Canada, July 2002.

[22] S. Zhong, and J. Ghosh, "A unified framework for model-based clustering", *Journal of Machine Learning Research,* Vol. 4, pp. 1001-1037, Dec. 2003.