



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

University of Wollongong
Research Online

University of Wollongong in Dubai - Papers

University of Wollongong in Dubai

2008

Applying and comparing hidden Markov model and fuzzy clustering algorithms to Web usage data for recommender systems

Shaghayegh Sahebi
University of Tehran, Iran

Farhad Oroumchian
University of Wollongong in Dubai, farhado@uow.edu.au

Ramtin Khosravi
University of Tehran

Publication Details

Sahebi, S., Oroumchian, F. & Khosravi, R. 2008, 'Applying and comparing hidden Markov model and fuzzy clustering algorithms to Web usage data for recommender systems', IADIS European Conference on Data Mining, 2008. Proceedings of, IADIS Press, Amsterdam, Netherlands, pp. 179-181.

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library:
research-pubs@uow.edu.au

APPLYING AND COMPARING HIDDEN MARKOV MODEL AND FUZZY CLUSTERING ALGORITHMS TO WEB USAGE DATA FOR RECOMMENDER SYSTEMS

Shaghayegh Sahebi

University of Tehran

N. Karegar Ave, P.O. Box 14395-515, Tehran, Iran,

Farhad Oroumchian

Wollongong University

Dubai, UAE

Ramtin Khosravi

University of Tehran

N. Karegar Ave, P.O. Box 14395-515, Tehran, Iran,

ABSTRACT

In this study, we apply and compare some of the methods of usage pattern discovery, like simple k-means clustering algorithm, fuzzy relational subtractive clustering algorithm, fuzzy mean field annealing (MFA) clustering and Hidden Markov Model (HMM), for recommender systems. We use metrics like prediction strength, hit ratio, precision, prediction ability and F-Score to compare the applied methods on the Web usage data. Fuzzy MFA and HMM acted better than other methods due to fuzzy nature of human behavior in navigation and extra information utilized in sequence analysis.

KEYWORDS

Web mining, usage pattern discovery, recommendation system, fuzzy clustering, Hidden Markov Model

1. INTRODUCTION

Web usage mining is the automatic discovery of user access patterns from Web servers and tries to discover valuable information from users' transactional data (Srivastava et al, 2000). A Web personalization system is defined as any system that tailors the Web experience for a particular user/a group of users (Mobasher et al, 2000). Many web mining techniques have been used in web personalization systems to discover usage patterns from Web data such as clustering techniques, association rule mining, and click pattern analysis. For example, Cadez et al. (2003) used the Expectation-Maximization (EM) algorithm on a mixture of Markov models for clustering user sessions. Joshi and Krishnapuram (2000) have used a fuzzy clustering approach for clustering user sessions.

In this study, we apply and compare various pattern recognition methods described in Section 2 to predict requested pages of users. The dataset is depicted in Section 3 and results are shown in Section 4.

2. APPLIED METHODS

We compared HMM (Bishop, 2006), a statistical model which the system modeling in it is considered as a Markov process, simple k-means clustering, RFSC (Suryavanshi et al, 2005), and Fuzzy MFA Clustering (Song et al, 2007) to Web usage data.

In k-means, we find the best cluster matches a session and sort the sum of user view time on this cluster pages to recommend the unvisited pages. For RFSC, we calculate the fuzzy membership matrix for a session;

then sort the clusters based on ascending order of membership degrees. For each cluster, the most important pages of it, calculated by a weighted sum of the viewing time of users, is recommended. In HMM, we select sequences of the length greater than L , and pad the first $L-1$ pages with all possible choices of pages. Viterbi algorithm (Forney, 1973) is used to derive most probable states and the probability for each sequence to recommend the most probable pages.

3. DATA AND MEASURES

We utilized the usage data of DePaul University computer science department which consisted of 13745 users' visiting duration on 683 pages. To solve the sparsity problem, which was a result of short sessions, we aggregated columns of the matrix using similar page URLs resulted in *aggregated data* and used principal component analysis (Bishop, 2006), a method for reducing data dimensions, on both main and aggregated data.

To evaluate the results, we applied some of the measures suggested in (Bose et al, 2006) and additional measures, taken from information retrieval literature, which were:

- Predictive Ability (PA): Percentage of hits with respect to the number of pages to be recommended.
- Prediction Strength (PS): Average number of recommendations made for a page.
- Hit Ratio (HR): Percentage of hits with respect to number of the sessions.
- Precision (Pr): Percentage of hits with respect to the number of recommendations for each session.
- F-Score (FS): A proportion of precision and predictive ability which is taken from Information Retrieval:

$$FScore = \frac{(PredictiveAbility \times Precision)}{(PredictiveAbility + Precision)}$$

4. EXPERIMENTAL RESULTS

To gain the proper number of clusters, which was 20 clusters, in k-means algorithm, we repeated it with different number of clusters on the main data using cosine similarity measure. The results are shown in Table 1.

In RFSC, a large distance matrix, with dimensions equal to the number of training data points, is kept. The dimensions of matrixes in MATLAB are restricted and less than and the training session number and we just could use 50 percent of our training data.

For fuzzy MFA algorithm we used 10 clusters. This algorithm does not use a straight approach to find the distance of data points, so we used the membership degree calculations of fuzzy k-means algorithm for the evaluation data.

Table 1. Evaluation of recommendations in different models

Model Name	PS	HR (%)	PA (%)	Pr (%)	F-Score (%)
k-means (20 Clusters)	12.34	-	42	6.8	11.7
k-means (10 Clusters)	12.48	-	44.2	7	12.1
RFSC (8 Clusters)	9.9	53	31	6	11
FMFA (10 Clusters)	9.98	65.5	43.15	7.8	14
HMM (20 Clusters)	11	60	38	11	18

Although the MFA algorithm takes a long time to converge, it is not as time-consuming as HMM. Due to high complexity of the training algorithm we bounded the algorithm to 50 iterations. We applied HMM with different L values for recommendations in length of k percent. It performed best for $L=4$ with 11 recommendations for each user.

According to the results of these algorithms in Table 1, HMM acts better than other algorithms, which is reasonable because of using the order of the viewed pages in conjunction with visited pages. The results of

RFSC algorithm were worse than simple k-means results. It may be due to the smaller number of clusters with respect to other algorithms. Albeit consuming lots of memory, the nonnecessity to define the number of clusters and obeying fuzzy partitioning condition which leads to less sensitivity to noises, and high speed are of advantages of this algorithm.

Besides, we can see that the results of fuzzy MFA clustering algorithm were better than other clustering algorithms. This also is a reasonable outcome because this algorithm provides a global optimum solution for clustering problem. By increasing the recommendation number to user, we can observe an increase in this algorithm's preciseness. Furthermore, we could enhance the results by applying the fuzzy MFA algorithm on different number of clusters and choosing the best of them.

5. CONCLUSION AND FUTURE WORK

In this study, we applied and compared clustering algorithms, like simple k-means, RFSC and fuzzy MFA clustering, and sequence analysis algorithm with HMM in Web usage-based recommender systems. The fuzzy MFA algorithm had been applied for the first time on Web usage data.

According to the experimental results section, HMM outperformed other algorithms but suffered from long learning time and after that Fuzzy MFA algorithm was better than other clustering algorithms.

For future works, we can improve these algorithms by using the cosine distance measure instead of Euclidean one in fuzzy MFA clustering algorithm, exploiting viewing time of users in HMM, and utilizing the domain knowledge like the structure or semantic information of Web site pages

REFERENCES

- Bishop C. M., 2006. *Pattern Recognition and Machine Learning*. Springer Publishers, Singapore.
- Bose, A. et al, 2006. Incorporating Concept Hierarchies into Usage Mining Based Recommendations. *Proceedings of WebKDD 2006*, Philadelphia, USA, pp 110-126.
- Cadez, I. et al, 2003. Model-based clustering and visualization of navigation patterns on a web site. *In Data Mining and Knowledge Discovery*, Vol. 7, No. 4, pp 399-424.
- Forney G. D., 1973. The Viterbi algorithm, *Proceedings of the IEEE*, pp 268-278.
- Joshi, A. and Krishnapuram, R., 2000. On mining web access logs. *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*. Dallas, USA, pp. 63-69.
- Mobasher, B. et al, 2002. Using sequential and non-sequential patterns for predictive web usage mining tasks. *Proceedings of the IEEE International Conference on Data Mining*. Maebashi City, Japan, pp. 669-672.
- Song, C. et al, 2007. A Mean Field Annealing Algorithm for Fuzzy Clustering. *Proceedings of Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. Haikou, China, pp. 193-197.
- Srivastava, J. et al, 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *In SIGKDD Explorations*, Vol. 1, No. 2, pp 12-23.
- Suryavanshi, B. S. et al, 2005. An Efficient Technique for Mining Usage Profiles using Relational Fuzzy Subtractive Clustering. *Proceedings of International Workshop on Challenges in Web Information Retrieval and Integration (WZH'05)*. Tokyo, Japan, pp. 23-29.