MINING USER ACCESS BEHAVIOR ON THE WWW

MEI-LING SHYU[1], SHU-CHING CHEN[2*], CHOOCHART HARUECHAIYASAK[1]

[1] Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33124-0640, USA
[2] Distributed Multimedia System Laboratory, School of Computer Science
Florida International University, Miami, FL 33199, USA

Abstract

In this paper, an affinity-based approach that provides good similarity measures for Web document clustering to discover user access behavior on the World Wide Web (WWW) is proposed. The proposed approach generates the similarity measures for groups of Web documents by considering the user access patterns. Any clustering algorithm using better similarity measures should yield better clusters for discovering user access behavior. By utilizing the discovered user access behavior, for example, the companies can previsely target their potential customers and convince them to purchase their products or services in electronic commerce. An experiment on a real data set is conducted and the experimental result shows that the proposed approach yields a better performance than the Cosine coefficient and the Euclidean distance method under the partitioning around medoid (PAM) method.

Keywords
Web document clustering, affinity-based, probabilistic model, user access behavior.

1 Introduction

Since its introduction in the early 1990s, the World Wide Web (WWW) has become an important means of providing and accessing information around the world. More and more information sources (for example, Web documents) have linked online through WWW, from personal data to scientific reports to up-to-the-minute satellite images. For example, in the context of electronic commerce, since many companies provide their product related or any usable information in the form of the *Uniform Resource Locator (URL)* pages on their Web sites for the customer convenience, the customer behavior can be captured by analyzing the user navigation through the company's Web site. For this purpose, Web usage mining, a process of applying data mining techniques to the discovery of usage patterns from Web data, has recently emerged as an analytical tool for management and decision-making [14].

One useful technique used in Web usage mining is the clustering algorithm. Clustering algorithms are used to create groups of similar documents to improve the efficiency and effectiveness of retrieval, or to determine the structure of the literature of the field [9]. For example, in electronic commerce, the companies can previsely target their potential customers and convince them to purchase their products or services by utilizing the discovered user access behavior from the Web document clustering results. According to some Web statistics surveyed in [8], the mean size for Web documents is *4.4 KB* with a median size of *2 KB* and has a maximum size of *1.6 MB*. The life span of Web documents is around *50* days. Due to these dynamic behaviors of the Web documents, clustering based on only the static quantities, i.e. terms or keywords, does not very well capture all characteristics of the Web documents. Hence, Web document clustering should incorporate some dynamic quantities such as the hyperlinks and the access patterns extracted from the user queries during the clustering process.

[*] This research was supported in part by NSF CDA-9711582.

Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference
Copyright © 2001

User access patterns can be extracted from the server log record [3, 7]. These usage patterns capture the dynamic quantities of the Web documents since each user browses through the Web site by following the hyperlinks provided in each Web document. To meet such a demand, we propose an affinity-based approach to assist in Web document clustering in this paper. The proposed approach uses an affinity-based probabilistic model to generate the similarity measures for the Web documents based on the user access patterns. In our previous work, the affinity-based probabilistic model has been applied to organize and manage multimedia databases [11, 12] and Web documents [13]. In [13], the clustering technique was applied in the document or *URL* level. In this paper, the proposed approach, however, applies the clustering process for groups of *URLs* belonging to one particular Web site. Consider the *URLs* on a software company Web site, these *URLs* can be categorized based on their high-level concepts or categories, such as product-related and customer service-related. Clustering analysis of the Web site might yield a result such as the programming development *URLs* are highly correlated to the software training *URLs*. The results from the clustering analysis can help managing the Web documents by reorganizing and customizing the company's Web site. For example, to help the customer navigating through the Web site, some hyperlinks of programming development *URLs* can be added into the *URLs* of software training, and vice versa. Other applications for the Web document clustering include Web search engine [2, 15], Web personalization [5] and adaptive Web site [6].

An experiment using a real data set [1] is conducted in this paper. In the experiment, three similarity/dissimilarity measure matrices obtained from Euclidean distance, cosine coefficient, and the proposed approach are applied to the partitioning around medoid (PAM) method. Any clustering algorithm with a good similarity measure approach should yield a low number of inter-cluster accesses since most of the related *URL* pages are grouped into the same cluster. The experimental result shows that the proposed approach performs better than the Euclidean distance and the cosine coefficient approaches since it yields the lowest number of inter-cluster accesses.

The paper is organized as follows. Next section describes the proposed affinity-based approach for similarity measures. In Section 3, an experiment on a real data set is conducted and the experimental result is presented. The paper is concluded in Section 4.

2 Affinity-Based Approach for Similarity Measures

The proposed approach uses an affinity-based probabilistic model that consists of a number of states (the *URL* groups) connected by transitions. The structure of the *URLs* in each group is modeled by the sequence of the states. The states are connected by directed arcs (transitions) that contain probabilistic and other data used to determine which state should be selected next.

2.1 The Affinity-Based Probabilistic Model

There are two probability distributions associated with each group – the state transition probability distribution ($\mathcal{A}$) and the initial state probability distribution (Π) in the affinity-based probabilistic model. The training data set consists of the user access patterns for the *URLs* in the groups in the Web site. Based on the user access patterns, the relative affinity values for pairs of the member *URLs* in one *URL* group are calculated. The variables to calculate the relative affinity value are defined as follows.

- $G = \{g_1, g_2, \ldots, g_g\}$ = a set of *URL* groups in the Web site
- n_i = number of *URLs* in each group g_i
- $Q = \{1, 2, \ldots, q\}$ = a set of instances (queries) in the training data set
- $use_{m,k}$ = usage pattern of *URL* m with respect to query k per time period

$$use_{m,k} = \begin{cases} 1 & \text{if } m \text{ is accessed by } k \\ 0 & \text{otherwise} \end{cases}$$

The relative affinity measures are used to indicate how frequently two *URLs* are accessed together. Two *URL* groups whose member *URLs* are accessed together more frequently are said to have a higher relative affinity relationship. Then, the relative affinity measures are used in the calculation of the entities of $\mathcal{A}$.

- $aff_{m,n}$ = affinity measure of *URL* m and *URL* n

$$aff_{m,n} = \sum_{k=1}^{q} use_{m,k} \times use_{n,k} \qquad (1)$$

- $f_{m,n}$ = the joint probability that refers to the fraction of the relative affinity of m and n in g_i with respect to the total relative affinity for all the *URLs* in g_i

$$f_{m,n} = \frac{aff_{m,n}}{\sum_{m \in g_i} \sum_{n \in g_i} aff_{m,n}} \qquad (2)$$

- f_m = the marginal probability

$$f_m = \sum_{n} f_{m,n} \qquad (3)$$

- $a_{m,n}$ = the conditional probability

$$a_{m,n} = \frac{f_{m,n}}{f_m} \qquad (4)$$

The conditional probability $a_{m,n}$ obtained from Equation 4 is the $(m,n)th$ entity of the state transition probability distribution $\mathcal{A}$ for each *URL* group.

The initial probability of a state (an *URL*) is the probability that the particular *URL* in a *URL* group can be the initial state for an incoming query. For any *URL* $m \in g_i$, the initial state probability is defined as the fraction of the number of occurrences of m with respect to the total number of occurrences for all the member *URLs* in g_i from the training data set. The initial state probability is defined as follows.

$$\Pi_i = \{\pi_{i,m}\} = \frac{\sum_{k=1}^{q} use_{m,k}}{\sum_{l=1}^{n_i} \sum_{k=1}^{q} use_{l,k}} \qquad (5)$$

Since the user access patterns of the *URLs* in each *URL* group are available from the training data set, the preference of the initial states (*URLs*) in each *URL* group can be obtained. The $\pi_{i,m}$ value is the probability that a state m in group g_i can be the initial state for an incoming query.

2.2 Similarity Measures

A similarity value $S(g_i, g_j)$ measures how well two *URL* groups g_i and g_j match the instances (queries) in the test data set.

$$S(g_i, g_j) = \sum_{O^k \in \mathcal{OS}} P(X, Y; g_i, g_j) F(N_k), \qquad (6)$$

where

- $N_k = k1 + k2$
- $\mathcal{OS}$ = a set of all the instance sets
- $O^k = \{o_1, \ldots, o_{N_k}\}$ = an instance set with the *URLs* belonging to g_i and g_j and generated by instance (query) k
- $X = \{x_1, \ldots, x_{k1}\}$ = a set of *URLs* belonging to g_i in O^k
- $Y = \{y_1, \ldots, y_{k2}\}$ = a set of *URLs* belonging to g_j in O^k
- $P(X, Y; g_i, g_j)$ = the joint probability of $X \in g_i$ and $Y \in g_j$

  $P(X, Y; g_i, g_j) = \prod_{u=2}^{k1} A_i(x_u \mid x_{u-1}) \Pi_i(x_1) \prod_{v=k1+2}^{N_k} A_j(y_{v-k1} \mid y_{v-k1-1}) \Pi_j(y_1)$, where A_i and Π_i are the state transition probability distribution and the initial state probability distribution for each *URL* group g_i, respectively.
- $F(N_k) = 10^{N_k}$.

  $F(N_k)$ is an adjusting factor since the number of the *URLs* in the instance set O^k accessed by query k is different.

3 An Experiment

An experiment is conducted to compare the quality of the similarity measures generated by the proposed approach with the similarity/dissimilarity measures generated by the Cosine coefficient and Euclidean distance approaches. A user browses through a Web site by either directly entering the *URL* location on the browser or by clicking on the hyperlinks provided within each *URL* page. This user access pattern can be obtained by extracting the information from the server log record or a particular Web site. A clustering strategy with a good similarity matrix should be able to cluster together the *URL* groups such that the requested *URL* pages from a user access pattern would fall into the same cluster as many as possible. Therefore, we use the number of inter-cluster accesses as the performance metric to compare the three similarity/dissimilarity matrices.

In the experiment, the similarity/dissimilarity matrices from the three approaches are generated based on the training data set. Then, these matrices are used to compare the the number of inter-cluster accesses based on the test data set under the partitioning around medoids (PAM) method.

3.1 Experimental Data Set

The experiment uses a real data set from Microsoft Web site (*Microsoft Anonymous Web Data*). It is available from University of California, Irvine's Knowledge Discovery in Databases (*UCI KDD*) Archive [1]. The data set was created by sampling and processing the *www.microsoft.com* logs. The data records the use of www.microsoft.com by approximately *38000* anonymous, randomly-selected users. These instances are divided into one training data set of *32711* instances and one test data set of *5000* instances. Each instance represents an individual user who is identified only by a sequential number or ID, for example, User *14988*, User *14989*, etc. The data set contains no personally identifiable information.

There are a total of *294 URLs* covered in the data set. From these *URLs*, we construct the attribute set of *39* items based on their concepts and contents, e.g., the attribute *country* is assigned for those *URLs* whose content are written in non-English languages and the attribute *programming* is assigned for the *URLs* whose contents are related to the programming languages. We then categorized these *URLs* into *13* groups based on these predefined attributes, e.g., *URL* group of *Networking and Server*, *URL* group of *Home, Education and Entertainment*, and *URL* group of *Service and Support*.

3.2 Similarity/Dissimilarity Measures

A variety of distance and similarity measures are used in document clustering process. Some well-known distance and similarity measures include the Euclidean distance, Manhattan (or city-block) distance, Dice coefficient, Jaccard coefficient, and Cosine coefficient [10]. For comparison purpose, the proposed approach and the Cosine coefficient for the similarity measures, and the Euclidean distance representing the dissimilarity measures are used in the experiments.

From the data set, we have *13* groups of *URLs* ($g_1, \ldots, g_{13}$) with each group contains some number of *URLs*. For the proposed approach, the state transition probability distributions and the initial state probability distributions are calculated using Equations 4 and 5. After these probability distributions are available, one similarity value for each pair of groups is generated. For Cosine coefficient and Euclidean distance approaches, since we have a set of *39* predefined attributes ($a_1, \ldots, a_{39}$), therefore, each *URL* can be represented by a binary vector of *39* dimensions. To represent a binary vector for a *URL* group, the centroid of the group can be calculated by averaging the attribute values of all *URLs* within the group. Once the attribute vectors for all **13** *URL* groups are calculated, the distances and the similarity matrices can be constructed as follows.

- The proposed approach – The similarity matrix is constructed using Equation 6.
- Cosine coefficient approach – The similarity matrix based in the cosine coefficient can be constructed using the following equation.

$$COSINE(g_i, g_j) = \frac{\sum_{k=1}^{39}(a_{i,k} \times a_{j,k})}{\sqrt{\sum_{k=1}^{39} a_{i,k}^2 \times \sum_{k=1}^{39} a_{j,k}^2}}.$$

- Euclidean distance approach – The dissimilarity matrix based on the Euclidean distance can be constructed using the following equation.

$$DISTANCE(g_i, g_j) = \sqrt{\sum_{k=1}^{39}(a_{i,k} - a_{j,k})^2}.$$

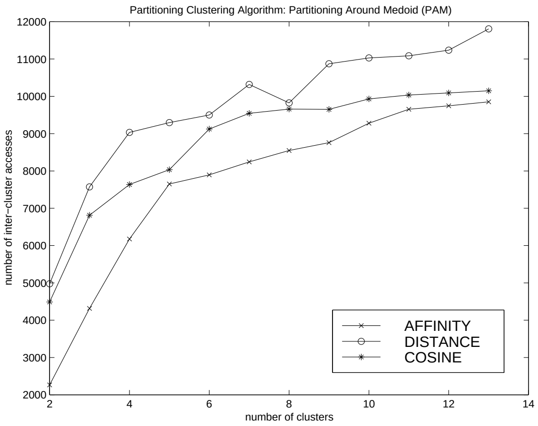

Figure 1: Partitioning Around Medoid (PAM) Method

3.3 The Partitioning Around Medoids (PAM) Method

For comparison purpose, the similarity/dissimilarity matrices generated from the proposed approach, the Cosine coefficient approach, and the Euclidean distance approach are applied to the partitioning around medoids (PAM) Method [4]. This method can be viewed as a variation of the well-known k-mean clustering algorithm. The idea of *PAM* clustering method is as follows. In order to obtain *k* clusters, the method selects *k* objects (which are called representative objects, or *k-medoid*) in the data set. The corresponding clusters are then found by assigning each remaining object to the nearest representative object based on the similarity or distance measures. The method differs from the k-mean algorithm in that once the representative objects are selected, they are fixed throughout the clustering process, whereas in k-mean method, the centroid of each cluster is recalculated when a new object is assigned to the cluster.

3.4 Performance Result

The result from using the Partitioning clustering method on three distance and similarity matrices is shown in Figure 1. Since we have a total of *13 URL* groups, we measure the performance by starting from the cluster size of *2* up to the cluster size of *13*. The result shows that the Euclidean distance (denoted by DISTANCE) gives a worse performance compared to the cosine coefficient (COSINE) while our affinity-based similarity measure (AFFINITY) yields the lowest number of inter-cluster accesses. In particular, when the number of clusters is from 2 to 4, our method yields a significant lower number of inter-cluster accesses. This is because that our proposed approach considers the user access patterns in the data set when determining the similarity between the Web documents. The closely related Web documents that have higher similarity relations are placed in the same cluster. When the number of clusters is small, the effectiveness of the clustering becomes more significant. Another observation is when the number of clusters increases, the number of inter-cluster accesses increases. The reason for the increase in the number of inter-cluster accesses with the increase in the number of clusters is that each cluster has only a few *URL* groups. This situation is the same for all the approaches. This explains why we have the similar amount of inter-cluster accesses, especially when the numbers of clusters are 10, 11, 12 and 13.

4 Conclusions

In this paper, an affinity-based probabilistic model to calculate the similarity measures between groups of Web documents based on the user access patterns for Web document clustering is presented. The proposed approach can capture the dynamic behavior of the Web documents since the user access patterns can be extracted from the Web server log record and represent the actual dynamic quantity of the user browsing through the Web site. The resulting Web document clusters can then better discover the user access behavior on the WWW.

To compare the quality of the similarity measures calculated by our approach against the other coefficients, we implemented and applied the partitioning around medoid (PAM) method. The number of inter-cluster accesses is used as the performance metric to compare performance among the different approaches. Any clustering algorithm with a good similarity coefficient should yield a low number of inter-cluster accesses since most of the related *URL* pages are grouped into the same cluster. The result shows that the proposed approach yields a better performance, i.e., a lower number of inter-cluster accesses, than the Euclidean distance and cosine coefficients.

References

[1] S.D. Bay, The UCI KDD Archive [http://kdd.ics.uci.edu]: Department of Information and Computer Science, University of California, Irvine, CA, 1999.

[2] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," *Proceedings of ACM SIGKDD International Conference*, pp. 407-415, 2000.

[3] R. Cooley, J. Srivastava, and B. Mobasher, "Data Preparation for Mining World Wide Web Browsing Patterns," *Journal of Knowledge and Information Systems*, Vol. 1, No. 1, pp. 5-32, 1999.

[4] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Inc., 1990.

[5] B. Mobasher, R. Cooley, J. Srivastava, "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs," *Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, November 1999.

[6] M. Perkowitz and O. Etzioni, "Adaptive Web Sites: Automatically Synthesizing Web Pages," *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1998.

[7] J. Pitkow, "In Search of Reliable Usage Data on the WWW," *The Sixth International World Wide Web Conference*, pp. 451-463, Santa Clara, CA, 1997.

[8] J. E. Pitkow, "Summary of WWW Characterizations," *The Seventh International World Wide Web Conference*, Brisbane, Australia, 1998.

[9] E. Rasmussen, "Chapter 16: Clustering Algorithms," in W. B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures & Algorithms*, pp. 419-442, Prentice Hall, 1992.

[10] G. Salton. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, 1989.

[11] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "A Probabilistic-Based Mechanism For Video Database Management Systems," *IEEE International Conference on Multimedia and Expo (ICME2000)*, pp. 467-470, New York City, USA, July 30-August 2, 2000.

[12] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Organizing a Network of Databases Using Probabilistic Reasoning," *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1990-1995, Nashville, Tennessee, USA, October 8-11, 2000.

[13] M.-L. Shyu, S.-C. Chen, and C.-M. Shu, "Affinity-Based Probabilistic Reasoning and Document Clustering on the WWW," *The 24th IEEE Computer Society International Computer Software and Applications Conference (COMPSAC)*, pp. 149-154, Taipei, Taiwan, October 25-27, 2000.

[14] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, Vol. 1, Issue 2, 2000.

[15] R. Weiss, B. Velez, M. A. Sheldon, C. Namprempre, P. Szilagyi, A.Duda, and D. K. Gifford, "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering," *Proceedings of the Seventh ACM Conference on Hypertext*, Washington, DC, March 1996.