Affinity-Based Similarity Measure for Web Document Clustering

Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
shyu@miami.edu

Shu-Ching Chen, Min Chen
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199, USA
{chens, mchen005}@cs.fiu.edu

Stuart H. Rubin
SPAWAR Systems Center (SSC) San Diego, Code 2734
53560 Hull Street
San Diego, CA 92152-5001, USA
stuart.rubin@navy.mil

Abstract

Compared to the regular documents, the major distinguishing characteristics of the Web documents is the dynamic hyper-structure. Thus, in addition to terms or keywords for regular document clustering, Web document clustering can incorporate some dynamic information such as the hyperlinks and the access patterns extracted from the user query logs. In this paper, we extend the concept of document clustering into Web document clustering by introducing the strategy of affinity-based similarity measure, which utilizes the user access patterns in determining the similarities among Web documents via a probabilistic model. Several comparison experiments are conducted using a real data set and the experimental results demonstrate that the proposed similarity measure outperforms the Cosine coefficient and the Euclidean distance method under different document clustering algorithms.

1. Introduction

In the area of document retrieval, many clustering algorithms have been introduced and extensively studied to improve the retrieval efficiency and effectiveness [8]. In particular, the similarity or dissimilarity measures, calculated based on the terms or keywords in the documents, are utilized in the clustering algorithms to group the similar documents.

Currently, there is an increasing need for effectively searching and managing the World Wide Web (WWW). In response to this need, a lot of research work has been carried out on Web document clustering. In most of the existing work, Web documents are clustered based on their static contents, such as keywords [14], terms, and phrases [13]. However, since WWW is a completely open environment, people can use any synonyms and/or abbreviations in their information sources, which potentially affects the effectiveness of those methods. More importantly, compared to the regular documents, the major distinguishing characteristics of the Web documents are the dynamic hyper-structure, i.e., each Web document can be modeled as an individual Uniform Resource Allocator (URL) that can be linked to or from other documents via the hyperlinks. As discussed in [4], the WWW can be regarded as a directed labeled graph, where the pages represent the nodes and links in the pages represent the edges of the graph. Thus, clustering based on the static content solely, such as terms or keywords, does not well capture the characteristics of the Web documents. Though the approaches in [4][7] tried to capture such dynamic aspects by taking into account the links among the Web documents, huge manual efforts were required to build the so-called link dictionaries. In addition, the links were still defined by the static contents.

In this paper, we propose a novel affinity-based similarity measure to assist Web document clustering. Different from the common methods which try to discover the similarities among the Web documents by the static contents or require huge manual efforts, our approach aims to develop an effective similarity measure based on users' preferences. More specifically, an affinity-based probabilistic model, which adopts the Markov model concept, is employed to automatically mine the document similarities based on user access patterns. Here, user access patterns can be extracted from the server log records which capture the dynamic characteristics of the Web documents in the sense that each user browsed through the Web site by following the hyperlinks provided in the Web documents. Intuitively, the more two Web documents are accessed together in user queries, the stronger the relation that exists between them. Such important information is then explored by the affinity-based probabilistic model to yield the similarity

measure, which can be easily plugged into the different clustering methods for the purpose of Web document clustering. In our previous work, the affinity-based model has been proposed and applied to the management of multimedia database [9][10] and Web documents [11]. Please note that in [11], the clustering technique was applied in the document or URL level; whereas in this work, we aim at exploring the similarity measures among the URL groups belonging to a particular web site, which is reasonable because normally the URLs on the web sites are naturally organized based on their high-level concepts. For instance, the URLs in a computer company might be categorized into company introduction, hardware-related information, programming development, software training, etc. Clustering analysis for the Web site might yield a result that the latter two groups of URLs are highly correlated, which can be used to facilitate the Web document management and Web site re-organization and customization. Web document clustering can also benefit the other applications including Web search engine [2], adaptive Web site [6], etc.

The rest of the paper is organized as follows. Section 2 describes our approach for calculating the similarity measure. A set of experiments is conducted on a real data set and the experimental results are presented in Section 3. Finally, the paper is concluded in Section 4.

2. Affinity-based similarity measure

The proposed affinity-based similarity measure is constructed by a probabilistic model, which adopts the Markov model concept. A Markov model is a well-researched mathematical construct which consists of a number of states connected by transitions. Thus, the mapping between the affinity-based probabilistic model and the construction of similarity measures for the URL groups are quite straightforward. That is, the states are utilized to model the URL groups, whereas the transitions represent the links among them. In order to obtain the similarity measure, the parameters for each URL group need to be formulated.

2.1. Formulations for parameters

In this study, each URL group is represented by a probabilistic model, where the structure of its URLs is modeled by the sequence of the states connected by directed arcs (transitions), which contain probabilistic and other data used to determine the state to be selected next.

Two probability distributions are associated with each group, which are described as follows.

1. $\mathcal{A}$ is the state transition probability distribution.

   In terms of the state transition probability in the probabilistic model, if two states (URLs) m and n are accessed together frequently, the probability that a traversal choice to state n given the current state is in m (or vice versa) should be higher. More specifically, it is used to indicate how closely two URLs are related.

2. Π is the initial state probability distribution.

   The initial probability distribution of a state (a URL) indicates the probability that the particular URL can be the initial state for an incoming query within a URL group. These initial probabilities indicate the preference of the URLs for the queries.

Both parameters are formulated based on a training data set.

2.1.1. Training data set. Suppose that, for a given web site S, there are N URL groups $G = \{g_1, g_2, \ldots, g_N\}$, where each of them contains n_i ($1 \le i \le N$) URLs. The training data set is constructed by collecting the user access patterns for all the URLs via a set of queries $Q = \{q_1, q_2, \ldots, q_q\}$ during the training process. Here, the user access pattern, denoted as $use_{m,k}$, is defined as

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

The user access patterns can be obtained by retrieving the access logs of the web site, where the more two URLs are accessed together, the more they are related to each other. Similarly, two URL groups are said to have a higher similarity value if their member URLs are accessed together more frequently.

2.1.2. State transition probability distribution. Based on the information in the training data set, the state transition probability distribution $\mathcal{A}$ for URL group g_i can be obtained via the following two steps.

1. The affinity measure of URLs m and n ($m, n \in g_i$) is defined as

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

2. The state transition probability distribution $\mathcal{A}$ is constructed by having $a_{m,n}$ as the element in the $(m,n)^{th}$ entry in $\mathcal{A}$, where

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

2.1.3. Initial state probability distribution. 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 member URLs in g_i from the training data set. The equation is given as follows.

$$\Pi_i = \{\pi_{i,m}\} = \frac{\sum_{k=1}^{q} use_{m,k}}{\sum_{l=1}^{n_i}\sum_{k=1}^{q} use_{l,k}} \quad (3)$$

The value of $\pi_{i,m}$ denotes the probability that a state (URL) m in group g_i can be the initial state for an incoming Web access.

2.2. Similarity measure

A similarity value measures how well two URL groups match the instances (queries) in the testing data set.

Let $X = \{x_1, x_2, \ldots, x_{k1}\}$ be a set of URLs belonging to the URL group g_i, $Y = \{y_1, y_2, \ldots, y_{k2}\}$ belonging to group g_j and $S(g_i, g_j)$ be the similarity measure between URL groups g_i and g_j. The similarity value $S(g_i, g_j)$ is calculated for each pair of URL groups g_i and g_j as shown in Equation (4), with the assumption that the sets X and Y are conditionally independent given g_i and g_j.

$$S(g_i, g_j) = \sum_{O^k \in OS} P(X, Y; g_i, g_j) F(N_k) \quad (4)$$

where OS is a set of all the instance sets, $N_k = k1+k2$ with $k1 = |X|$, $k2 = |Y|$, and $O^k = \{o_1, o_2, \ldots, o_{N_k}\}$ is an instance set with the URLs belonging to g_i and g_j and generated by instance (query) q_k.

Since the sets X and Y are conditionally independent given g_i and g_j, it can be further defined as:

$$P(X, Y; g_i, g_j) = P(X; g_i) P(Y; g_j) \quad (5)$$

$$P(X; g_i) = \Pi_{u=2}^{k1} A_i(x_u \mid x_{u-1}) \Pi_i(x_1) \quad (6)$$

$$P(Y; g_j) = \Pi_{v=k1+2}^{N_k} A_j(y_{v-k1} \mid y_{v-k1-1}) \Pi_j(y_1) \quad (7)$$

where $A_i(x_u|x_{u-1})$ and $\Pi_i(x_1)$ correspond to a_{x_{u-1},x_u} in $\mathcal{A}$, and π_{i,x_1} in Π_i, respectively, for g_i. As can be seen from the above equations, the similarity measure between g_i and g_j is calculated using the parameters $\mathcal{A}_i$, Π_i, $\mathcal{A}_j$, Π_j, respectively. $F(N_k) = 10^{N_k}$ is an adjusting factor since the number of URLs in the instance set O^k accessed by query q_k may be variable.

The resulting similarity values can be constructed as a matrix of N by N, where N is the number of URL groups in the Web site. The similarity matrix is symmetric, which means $S(g_i, g_j)$ is equal to $S(g_j, g_i)$. In the next section, we utilize our approach described in this section to construct the similarity matrix on a real data set and compare its performance to two other similarity and dissimilarity measures (for short, *clustering measures*) using different clustering methods.

3. Experiments and results

In the experiments, we would like to demonstrate the fact that the more effective a *clustering measure* is, the better clustering result could be achieved using the same clustering algorithm. In this section, four experiments are conducted to justify the effectiveness of our proposed similarity measure for Web document clustering. The three *clustering measures* from the Euclidean distance, Cosine coefficient, and our proposed affinity-based approach are used in four clustering methods for comparison. In addition, the number of inter-cluster accesses is used as the performance metric in the experiments.

3.1. The experimental data set

A real data set called the Microsoft Anonymous Web Data, which belongs to Microsoft Web site, is used to construct the training data set and the testing data set for our experiments. It was obtained from University of California, Irvine's Knowledge Discovery in Databases (UCI KDD) Archive [1] and consisted of 294 URLs and approximately 38,000 randomly-selected anonymous user accesses.

From these URLs, we construct the attribute set of 39 items based on their concepts and contents. For example, the attribute *programming* is assigned to the URLs whose content 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 Service and Support, etc. The set of user accesses is randomly divided into two data sets: training data set and testing data set, which contains 32,711 and 5,000 instances, respectively. Here, the training data set is applied to construct the three different *clustering measures*, and testing data set is utilized to test the clustering performance with these measures.

3.2. Similarity matrix for the proposed approach

Using the training data set, the state transition probability distributions and the initial state probability distribution for each URL group can be obtained according to Equations (1) to (3) given in Section 2.1. Then, the similarity matrix is constructed by following the steps described in Section 2.2.

3.3. Other clustering measures for comparison

A variety of *clustering measures* are used in document clustering process. Among them, the Euclidean distance, Manhattan (or city-block) distance, Dice coefficient, Jaccard coefficient, and Cosine coefficient are the most well-known measures. For comparison purposes, we select two measures, i.e., Euclidean distance (a dissimilarity measure) and Cosine coefficient (a similarity measure). Further discussions concerning the usage of different coefficients in document retrieval can be found in [12].

As mentioned earlier, we have a set of 39 predefined attributes and 13 URL groups, which contain various numbers of URLs. Therefore, each URL is represented by a binary vector of 39 dimensions and the URL group is represented by the centroid of the group, which is calculated by averaging the attribute values of all the URLs within the group. Let, $\{a_{i,k} \mid 1 \le k \le 39\}$ and $\{a_{j,k} \mid 1 \le k \le 39\}$ be two attribute vectors for URL groups g_i and g_j, respectively; the *clustering measures* between them can be constructed as follows.

- Euclidean distance approach

The dissimilarity value based on the Euclidean distance can be constructed using the following equation.

$$Dis(g_i, g_j) = \sqrt{\sum_{k=1}^{39} (a_{i,k} - a_{j,k})^2} \tag{8}$$

- Cosine coefficient approach

The similarity value based on the Cosine coefficient can be obtained using the following equation.

$$Cos(g_i, g_j) = \frac{\sum_{k=1}^{39} (a_{i,k} \times a_{j,k})}{\sqrt{\sum_{k=1}^{39} a_{i,k}^2 \times \sum_{k=1}^{39} a_{j,k}^2}} \tag{9}$$

3.4. Clustering algorithms for comparison

Two different classes of clustering algorithms have been widely used in document clustering, namely, nonhierarchic and hierarchic methods. The former approach partitions the document into disjoint groups, whereas the latter one organizes the documents into a hierarchical or treelike structure. For comparison purpose, four clustering algorithms, namely the Partitioning Around Medoids (PAM) method, Single-Link method, Group Average Link method, and Complete Link method, are implemented to apply upon the three different *clustering measures*. Among these four algorithms, the first one belongs to the nonhierarchic method; whereas, the others are hierarchic methods. Their algorithms are briefly introduced below, while the detailed discussions can be found in [2][3][5].

- Partitioning Around Medoids (PAM) method

The PAM method can be regarded as a variation of the well-known *k*-means clustering algorithm. The major difference is that for the PAM method, 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.

- Three hierarchic methods

The algorithms for these three hierarchic methods are defined similarly in two steps as follows:

1. Each item, which needs to be clustered, forms a singleton cluster.
2. While there is more than one cluster, the clusters with the maximum similarity are merged (ties are broken arbitrarily) and the similarity between the newly merged cluster and the remaining clusters is recomputed.

The resulting clustered collection consists of a hierarchy of items in which small clusters with very strongly related items are nested within larger clusters with less strongly related items. The difference among these methods lies in the way that the similarity between nonsingleton clusters is defined [3].

These four algorithms are applied upon the three *clustering measures*, obtained from the training data set, to produce the clustering results. In the next subsection, the performance metric and the comparison results achieved using the testing data set are presented.

3.5. Performance metric and results

As discussed in the previous subsection, four clustering algorithms are implemented and applied to perform the clustering process. Then, the testing data set with 5,000 instances is used for the purpose of performance comparison. A good clustering result should ensure that the requested URL pages in the same user access pattern 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 *clustering measures*. In other words, the lower the number of inter-cluster accesses is, the better the performance achieved.

Figure 1 shows the results obtained by applying the PAM clustering method to the three *clustering measures*. Since we have totally 13 URL groups, the performances are evaluated with the number of clusters from 1 (i.e., all the URL groups are contained in one cluster) to 13 (i.e., 1 URL group per cluster). The results demonstrate that our affinity-based similarity measure (denoted by AFFINITY) produces the best performance in terms of

yielding the lowest number of inter-cluster accesses; whereas in most cases, the Euclidean distance (DISTANCE) gives the worst performance, followed by the cosine coefficient (COSINE). In particular, when the number of clusters is set within a reasonable range, in this case, 3 to 11 (i.e., it is not extremely small or big), the effectiveness of our method becomes more significant in the sense that a dramatic lower number of inter-cluster accesses is produced. This proves that by considering the user access patterns in the construction of affinity-based similarity measure, the dynamic hyper-structure of Web documents is explored effectively. Thus, during the clustering process, the closely related URL groups (Web documents) are placed in the same cluster. Another observation is that the number of inter-cluster accesses increases as the number of clusters increases, which is because with more clusters, the number of URL groups contained in each cluster decreases and the probability of accessing the Web documents belonging to different clusters becomes higher.

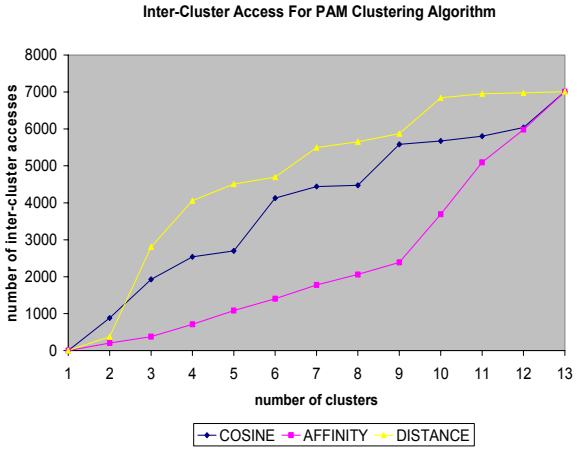
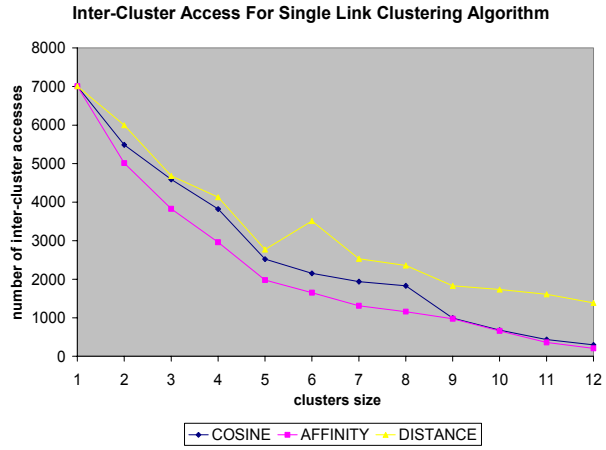


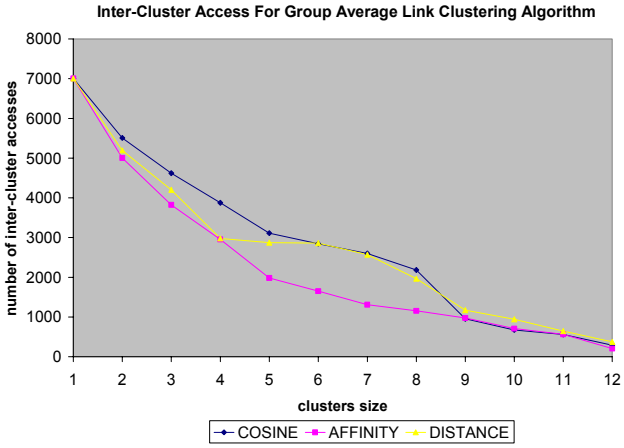
Figure 1. Partition Around Medoid (PAM) method

Figures 2 (a) to (c) demonstrate the results produced by the three hierarchic clustering methods, namely the Single Link method, Group Average Link method, and Complete Link method, respectively. Note that since the parameter *cluster size*, which represents the maximal number of URL groups can be contained in each cluster, needs to be pre-defined for the hierarchic clustering methods to perform, in the following three experiments, "*cluster size*" instead of "*number of clusters*" is set from 1 to 13 for performance evaluations. As can be seen from Figure 2, our proposed affinity-based similarity measure yields the best results for all the three hierarchic clustering methods, especially when the *cluster size* is set to a reasonable range.

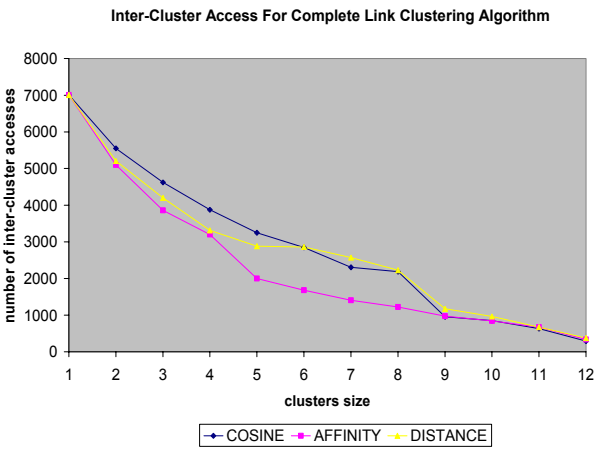
As for all these different clustering methods, the proposed similarity measure can achieve the best performance, which justifies its effectiveness for Web document clustering.



(a) Single Link Clustering method



(b) Group Average Link method



(c) Complete Link method

Figure 2. The hierarchic clustering methods

4. Conclusions

In this paper, a novel affinity-based similarity measure is proposed for Web document clustering. In particular, in order to capture the dynamic characteristics of the Web documents, an affinity-based probabilistic model, which adopts the Markov model concept, is employed to automatically explore the similarity measure between a pair of URL groups based on the user access patterns. The experimental results exemplify the effectiveness of our proposed similarity measure approach in the sense that it produces much better clustering results than the other two clustering measures, which utilize only the static information in the data under various clustering methods.

5. Acknowledgement

For Mei-Ling Shyu, this research was supported in part by NSF ITR (Medium) IIS-0325260. For Shu-Ching Chen, this research was supported in part by NSF EIA-0220562 and HRD-0317692. For Stuart Rubin, this research was supported by SSC, code 270.

6. References

[1] S. D. Bay, The UCI KDD Archive. http://kdd.ics.uci.edu.

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

[3] M. R. Ellen, "Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval," *Information Processing & Management*, Vol. 22, No. 6, 1986, pp. 465-476.

[4] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis, "THESUS: Organizing Web Document Collections Based on Link Semantics," *The VLDB Journal*, Vol. 12, No. 4, 2003, pp. 320-332.

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

[6] B. Mobasher, et al., "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs," *Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, 1999.

[7] S. Modha, and W. Spangler, "Clustering Hypertext with Applications to Web Searching," *Proceedings of the Conference on Hypertext*, 2000, pp. 143-152.

[8] E. Rasmussen, "Chapter 16: Clustering Algorithms," *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, 1992, pp. 419-442.

[9] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "A Probabilistic-Based Mechanism for Video Database Management Systems," *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME2000)*, 2000, pp. 467-470.

[10] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Organizing a Network of Databases Using Probabilistic Reasoning," *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 2000, pp.1990-1995.

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

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

[13] O. Zamir and O. Etzioni, "Web Document Clustering: A Feasibility Demonstration," *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 46-54.

[14] O. Zamir, O. Etzioni, O. Madani, and R. Karp, "Fast and Intuitive Clustering of Web Documents," *Proceedings of the ACM SIGMOD International Workshop on Data Mining and Knowledge Discovery*, 1997, pp. 287-290.