Affinity-Based Probabilistic Reasoning and Document Clustering on the WWW

Mei-Ling Shyu
Department of Electrical and Computer Engineering,
University of Miami,
Coral Gables, FL 33124-0640, USA
shyu@miami.edu

Shu-Ching Chen
School of Computer Science,
Florida International University,
Miami, FL 33199, USA
chens@cs.fiu.edu

Chi-Min Shu
Department of Environmental Safety Engineering,
National Yunlin University of Science and Technology,
Yunlin, Taiwan, R.O.C.
shucm@pine.yuntech.edu.tw

Abstract

The World Wide Web (WWW) has become one of the fastest growing applications on the Internet today. More and more information sources have linked online through WWW, but finding information on the WWW is also a great challenge. For most of the users, the information retrieved is not well organized and the access time is considered high on the WWW currently. Therefore, there is a need to develop a good mechanism to organize and manage the tremendous size and various kinds of information to facilitate the functionality of a search engine for information retrieval on the WWW. In response to such a demand, we propose a Markov Model Mediator (MMM) mechanism which employs the affinity-based data mining techniques to organize and manage the information sources so that the most relevant documents are clustered together to achieve higher recall and precision values for information retrieval on the WWW.

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. For those information providers, they simply put the information on the Web servers. For the users, they can access information by requesting the servers to send the information via the Web browsers. Though the WWW provides such a convenient way for putting and getting information, the information retrieved is not well organized and the access time is considered high on the WWW for most of the users.

With the increasing number of information sources on the WWW, the need to develop a good mechanism to organize and manage the tremendous size and various kinds of information for information retrieval becomes important. One technique to search information from the WWW is by keyword-based querying. Keyword searching has been an immediate and efficient way to specify and find related information that the user inquires. However, since the WWW is a completely open environment, people can use any synonyms and/or abbreviations in their information sources. Potentially, a large amount of information is retrieved using keyword matching, but the retrieval precision is low due to inappropriate match of keywords. In addition, the recent increase in popularity of the WWW has led to a considerable increase in the amount of traffic over the Internet which makes finding information on the WWW time-consuming.

Imagine that given a query, a Web search engine produces hundreds of hits that represent documents supposedly relevant to the query and only a small portion of the located documents is actually relevant to the query. It takes so much time for the user to wait for the browser to display such a huge list of documents and to browse through the list. If those documents can be organized into clusters with respect to their degrees of relevance to the issued queries, the time to browse through the documents for the user can be reduced significantly since higher recall and precision values for information retrieval can be achieved. Towards this demand, the data mining techniques which can discover qualitative and quantitative patterns from the query usage patterns on the WWW can be beneficial.

Data mining is a process to extract nontrivial, implicit, previously unknown and potentially useful information from data. Data mining involves data analysis techniques which develop methods for extracting valuable knowledge from the huge repositories of data – most of which will remain unseen by humans. Three of the most common methods to mine data are association rules [12] [13], data classification [3] [8] and data clustering [5] [16]. Association rules discover the co-occurrence associations among data. Data classification is the process that classifies a set of data into different classes according to some common properties and classification models. Finally, data clustering groups physical or abstract objects into disjoint sets that are similar in some respect.

Many data clustering strategies have been proposed in the literature. Methods that rely on the designers to give hints on what objects are related require the domain knowledge of the designers [2] [9]. Syntactic methods such as depth first and breadth first, determine a clustering strategy based solely on the static structure of the database [7]. The disadvantages of this strategy are that it ignores the actual access patterns and the queries might not traverse the database according to the static structure. The third type of methods gather the statistics of the access patterns and partition the objects based on the statistics [14]. Other strategies such as the placement tree clustering method in [1] and the decomposition-based simulated annealing clustering method [6] combine two or all of the above strategies.

In this paper, the *Markov Model Mediator (MMM)* mechanism is proposed to organize and manage the documents for information retrieval on the WWW. The large number of documents is represented as a browsing graph with each document represented as a node in the browsing graph. The connectivity of the nodes in the browsing graph is determined by the structural relationships between two documents. Two nodes are connected in the browsing graph only if the two corresponding documents have structurally equivalent terms. A set of data, i.e., the usage patterns and access frequencies of the queries issued on the WWW, together with the structure of the document are used to generate the training traces for the proposed affinity-based data mining process. The MMM mechanism employs the affinity-based data mining techniques – document clustering and probabilistic reasoning. By analyzing the statistics of the query usage patterns from a set of historical data, probabilistic reasoning derives sets of probability distributions for the MMMs. The probability distributions are required in the proposed stochastic process which leads to document clustering. Document clustering groups the tremendous amount of documents into a set of clusters with respect to their degrees of relevance to the queries. The idea is to organize the documents such that the user can get the closely related links for the query in the minimal amount of time. The *MMM* mechanism can be incorporated into a Web search engine to facilitate the functionality of the search engine. Instead of producing and displaying a huge list of unorganized documents to the user, the search engine can display the documents in clusters. Documents in the same cluster are potentially closely related to a certain application domain. In addition, since queries tend to access information from the closely related documents, document clustering can improve recall by effectively matching queries against clusters and retrieving clusters with the highest similarity measure.

The organization of this paper is as follows. Section 2 briefly overviews the MMM mechanism. In section 3, the affinity-based data mining process which includes probabilistic reasoning and a stochastic process for document clustering is introduced. Conclusions are presented in section 4.

2 Markov Model Mediator (MMM) Mechanism

The *Markov Model Mediator (MMM)* mechanism adopts both the *Markov Model* framework and the *mediator* concept. A Markov model is a well-researched mathematical construct which consists of a number of states connected by transitions. A mediator is defined as a program that collects information from one or more sources, processes and combines it, and exports the resulting information [15]. In our previous study, we proposed the use of the MMM mechanism to facilitate the functionality of a *multimedia database management system (MMDBMS)* [10, 11]. In this paper, the functions of the MMM mechanism are extended to organize the documents into clusters with respect to the degrees of relevance for efficient information retrieval on the WWW.

Each document on the WWW is modeled by an MMM. A document can be a web page for a company which sells computer hardware. Each document consists of a set of terms which are the components of a computer such as a terminal, a mouse, and a keyboard. Each term has a set of attributes like price, size, color, etc. An MMM is represented by a 6-tuple $\lambda = (\mathcal{S}, \mathcal{F}, \mathcal{A}, \mathcal{B}, \Pi, \Psi)$ where $\mathcal{S}$ is a set of terms called states; $\mathcal{F}$ is a set of attributes; $\mathcal{A}$ is the state transition probability distribution; $\mathcal{B}$ is the observation symbol probability distribution; Π is the initial state distribution; and Ψ is a set of multimedia augmented transition networks (ATNs). The multimedia ATN is based on the augmented transition network (ATN) model and is used as a semantic model for multimedia presentations, multimedia database searching, and multimedia browsing [4].

The structure of the terms (in $\mathcal{S}$) in a document is modeled by the sequence of the MMM states. Each term has its own set of attributes (in $\mathcal{F}$). The states are connected by directed arcs (transitions) which contain probabilistic and other data used to determine which state should be selected next. $\mathcal{A}, \mathcal{B}$, and Π are the probability distributions for an

Table 1. The query usage patterns.

$use_{k,m}$	states (terms)												
	1	2	1	2	3	1	2	3	4	5	1	2	3
	$t_{1,1}$	$t_{1,2}$	$t_{2,1}$	$t_{2,2}$	$t_{2,3}$	$t_{3,1}$	$t_{3,2}$	$t_{3,3}$	$t_{3,4}$	$t_{3,5}$	$t_{4,1}$	$t_{4,2}$	$t_{4,3}$
q_1	1	0	1	1	1	0	0	0	0	0	0	1	0
q_2	1	1	1	0	0	1	1	1	1	1	1	0	1
q_3	0	1	1	0	1	1	1	1	1	0	0	0	0
q_4	1	0	0	1	1	0	1	0	0	0	1	1	0
q_5	1	1	1	1	0	1	0	1	0	1	1	0	1
q_6	1	0	1	0	0	1	0	1	1	1	0	1	0
q_7	0	1	0	1	1	0	1	1	0	1	1	0	1
q_8	0	0	1	1	0	0	1	0	1	1	0	0	1
q_9	0	1	0	1	1	0	0	1	1	0	0	1	1
q_{10}	1	0	0	1	1	1	0	1	0	0	1	0	1

MMM and will be mined in the affinity-based data mining process.

3 Affinity-Based Data Mining Process

3.1 Relative Affinity Measures

We use the relative affinity measurements to indicate how frequently two terms are accessed together. Two documents whose terms are accessed together more frequently are said to have a higher relative affinity relationship. Realistically, the applications cannot be expected to specify these affinity values. Therefore, formulas to calculate these relative affinity values need to be defined.

We will use the following simple document system as an example to illustrate how the MMM mechanism together with the affinity-based data mining techniques are used for document clustering. Assume there are four documents d_1, d_2, d_3, and d_4. Each document has its own set of terms with each term denoted by $t_{i,j}$, where i represents the ith document (i.e., d_i) and j represents the jth term in that document. For example, d_1 has two terms that are denoted by $t_{1,1}$ and $t_{1,2}$. Let the total number of distinct attributes in the system be **22**. Moreover, assume a set of query usage patterns and access frequencies with ten queries is used to generate the training traces. Table 1 shows the usage patterns of the terms versus the sample queries. If a term was accessed by a certain query, then the corresponding entity has a value **1**. For example, the term $t_{1,2}$ has been accessed by queries q_2, q_3, q_5, q_7, and q_9. The access frequencies of the sample queries are shown in Table 2. For example, the access frequency of query q_2 is 50.

Let Q = $\{1, 2, \ldots, 10\}$ be the set of sample queries that ran on the documents $d_1, d_2, \ldots, d_4$ with the term set OC = $\{$1, 2, . . ., 13$\}$ in the system. Define the variables:

- n_i = number of terms in document d_i
- $use_{m,k}$ = usage pattern of term m with respect to query q_k per time period

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

- $access_k$ = access frequency of query q_k per time period
- $aff_{m,n}$ = affinity measure of terms m and n

$$aff_{m,n} = \sum_{k=1}^{10} use_{m,k} \times use_{n,k} \times access_k \qquad (1)$$

Table 2. The query access frequencies.

$access_k$	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}
access frequency	35	50	30	60	80	45	20	30	10	40

3.2 Probabilistic Reasoning Method

3.2.1 State Transition Probability Distribution

Two documents whose terms are accessed together more frequently are said to have a higher relative affinity relationship. Accordingly, in terms of the state transition probability in a Markov Chain, if two documents have a higher relative affinity relationship, the probability that a traversal choice to node n given the current node is in m (or vice versa) should be higher. Therefore, the conditional probability $a_{m,n}$ is the $(m, n)th$ entity of the state transition probability distribution $\mathcal{A}$. Define

- $f_{m,n} = \frac{aff_{m,n}}{\sum_{m \in d_i} \sum_{n \in d_i} aff_{m,n}}$ = the joint probability which refers to the fraction of the relative affinity of terms m and n in a document with respect to the total relative affinity for all the terms in a document
- $f_m = \sum_n f_{m,n}$ = the marginal probability
- $a_{m,n} = \frac{f_{m,n}}{f_m}$= the conditional probability which refers to the state transition probability for an MMM

Table 3 gives the constructed state transition probability distribution for d_1.

Table 3. The state transition probability distribution $\mathcal{A}$ for d_1. For example, the transition goes from state 1 (term $t_{1,1}$) to state 2 (term $t_{1,2}$) is 0.2955.

state	1	2
1	0.7045	0.2955
2	0.4062	0.5938

3.2.2 Observation Symbol Probability Distribution

The observation symbol probability denotes the probability of observing an output symbol from a state. Here, the observed output symbols represent the attributes and the states represent the terms. Since a term has one or more attributes and an attribute can appear in multiple terms, the observation symbol probabilities shows the probabilities an attribute is observed from a set of terms.

A temporary matrix BB whose columns are the terms in a document and rows are all the distinct attributes in the environment is assigned a value **1** or **0** for each entity of BB. It indicates whether an attribute appears in a term of the document.

$$BB_{s,t} = \begin{cases} 1 & \text{if attribute } s \text{ appears in term } t \\ 0 & \text{otherwise} \end{cases}$$

After BB is constructed, the observation symbol probability distribution $\mathcal{B}$ can be obtained via normalizing BB per column. In other words, the sum of the probabilities which the attributes are observed from a given term should be 1.

Table 4 is the constructed observation symbol probability distribution for d_1. From this table, we can obtain information such as the probability that attribute 3 is observed given the current state is 1 ($t_{1,1}$) in d_1 is 1/4.

3.2.3 Initial State Probability Distribution

Since the information from the training traces is available, the preference of the initial states for queries can be obtained. For any term $m \in d_i$ (the ith document), the initial state probability is defined as the fraction of the number of occurrences of term m with respect to the total number of occurrences for all the member terms in document d_i from the training traces.

Table 4. $\mathcal{B}$ for d_1.

	1	2
1	1/4	0
2	1/4	0
3	1/4	0
4	1/4	1/4
5	0	1/4
6	0	1/4
7	0	1/4
8	0	0
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0
21	0	0
22	0	0

$$\Pi_i = \{\pi_m\} = \frac{\sum_{k=1}^{10} use_{m,k}}{\sum_{l=1}^{n_i} \sum_{k=1}^{10} use_{l,k}} \qquad (2)$$

Hence, the four initial state probability distributions for documents d_1 to d_4 can be determined by using Equation 2. For example, Π_1 = $[\frac{6}{11}\ \frac{5}{11}]$ for d_1.

3.3 Similarity Measure of Two Documents – A Stochastic Process

A similarity value measures how well two documents together match the observations generated by the sample queries. The similarity value is formulated under the assumptions that the observation set O^k is conditionally independent given X and Y, and the sets $X \in d_i$ and $Y \in d_j$ are conditionally independent given d_i and d_j. Let $N_k = k1 + k2$, $\mathcal{OS}$ be a set of all observation sets, and $S(d_i, d_j)$ be the similarity measure between documents d_i and d_j. The similarity values are computed for the pairs of documents that are connected in the browsing graph.

$$S(d_i, d_j)$$
$$= \sum_{O^k \in \mathcal{OS}} P(O^k \mid X, Y; d_i, d_j) P(X, Y; d_i, d_j) F(N_k),$$

where

- $O^k = \{o_1, \ldots, o_{N_k}\}$ is an observation set with the attributes belonging to d_i and d_j and generated by query q_k

- $X = \{x_1, \ldots, x_{k1}\}$ is a set of terms belonging to d_i in O^k
- $Y = \{y_1, \ldots, y_{k2}\}$ is a set of terms belonging to d_j in O^k
- A_i, B_i, and Π_i are the state transition probability distribution, the observation symbol probability distribution, and the initial state probability distribution for document d_i, respectively. These probability distributions are formulated using the proposed probabilistic reasoning method.
- $P(O^k \mid X, Y; d_i, d_j)$ = the probability of occurrence of O^k given $X \in d_i$ and $Y \in d_j$ = $\prod_{u=1}^{k1} B_i(o_u \mid x_u)$ $\prod_{v=k1+1}^{N_k} B_j(o_v \mid y_{v-k1})$
- $P(X, Y; d_i, d_j)$ = the joint probability of $X \in d_i$ and $Y \in d_j = \prod_{u=2}^{k1} A_i(x_u \mid x_{u-1})\ \Pi_i(x_1) \prod_{v=k1+2}^{N_k}$ $A_j(y_{v-k1} \mid y_{v-k1-1})\ \Pi_j(y_1)$
- $F(N_k)$ = an adjusting factor which is used because the lengths of the observation sets are variable = 10^{N_k}

Table 5. The similarity values for pairs of documents. For example, the similarity value between d_1 and d_4 is 20.23.

$S(d_i, d_j)$	d_1	d_2	d_3	d_4
d_1	-	18.93	0	20.23
d_2	18.93	-	7.28	0
d_3	0	7.28	-	6.95
d_4	20.23	0	6.95	-

The resulting similarity values for the documents are shown in Table 5. As can be seen from Table 5, the similarity values are symmetric which means $S(d_i, d_j)$ is equal to $S(d_j, d_i)$. When two documents have no structurally equivalent terms, the similarity value between them is **0** because there is no connectivity between these two nodes (documents) in the browsing graph.

3.4 Document Clustering Strategy

The similarity values calculated from the stochastic process are transformed into the branch probabilities among the nodes (documents) in the browsing graph. Then the stationary probability ϕ_i for each node i in the browsing graph can be obtained from the branch probabilities. The stationary probability ϕ_i denotes the relative frequency of accessing node i (document d_i) in the long run.

$$\sum_i \phi_i = 1 \qquad\qquad \phi_j = \sum_i \phi_i P_{i,j} \quad j = 1, 2, \cdots \qquad (3)$$

Table 6. The branch probabilities transformed from the similarity values (in Table 5). For example, the branch probability from d_1 to d_4 is 0.5166.

$P_{i,j}$	d_1	d_2	d_3	d_4
d_1	-	0.4834	0	0.5166
d_2	0.7222	-	0.2778	0
d_3	0	0.5116	-	0.4884
d_4	0.7443	0	0.2557	-

The similarity values in Table 5 are then transformed into the branch probability $P_{i,j}$ for nodes i and j (as shown in Table 6) for the browsing graph. The transformation is executed by normalizing the similarity values per row to indicate the branch probabilities from a specific node (document) to all its accessible nodes (documents). For example, d_1 has similarity value **18.93** with d_2 and **20.23** with d_3, and these two similarity values are transformed into the branch probabilities **0.4834** and **0.5166**. After the branch probabilities are obtained, the stationary probabilities for the four documents can be computed by using Equation 3. Table 7 lists the stationary probabilities for the four documents.

Table 7. The stationary probability.

document	d_1	d_2	d_3	d_4
ϕ	0.3667	0.2455	0.1333	0.2545

Our document clustering strategy is traversal based and greedy. Documents are partitioned with the order of their stationary probabilities. The document which has the largest stationary probability is selected to start a document cluster. While there is room in the current document cluster, all documents accessible in terms of the browsing graph from the current member documents of the document cluster are considered. The document which has the largest stationary probability is selected and the process continues until the document cluster fills up. At this point, the next un-partitioned document from the sorted list starts a new document cluster, and the whole process is repeated until no un-partitioned documents remain. In this example, if the number of documents per cluster is two, then there will be two clusters. The first cluster consists of documents d_1 and d_4; while the second cluster has documents d_2 and d_3. The time complexity for this document clustering strategy is $O(d \log d)$, where d is the number of documents.

4 Conclusions

In this paper, we introduced a mathematically sound framework, Markov model mediator (MMM) mechanism, to facilitate the functionality of a Web search engine for fast information retrieval on the WWW. The MMM mechanism employs the proposed affinity-based data mining process that includes probabilistic reasoning and document clustering. The probabilistic reasoning derives the sets of probability distributions for the MMMs and are used in the proposed stochastic process for document clustering. A document clustering strategy based on the MMM mechanism is proposed to partition the documents into a set of beneficial document clusters. Since a document cluster consists of several related documents which are usually required for queries in the same application domain, document clustering can improve recall by effectively matching queries against clusters. More experiments to compare our proposed MMM mechanism with other clustering approaches will be conducted in the near future.

References

[1] V. Benzaken and C. Delobel, “Enhancing performance in a persistent object store: Clustering strategies in O_2,” in A. Dearle, G.M. Shaw, and S.B. Zdonik, editors, Implementing Persistent Object Bases: Principles and Practice, pp. 403-412, Morgan Kaufmann, 1991.

[2] M.J. Carey, D.J. DeWitt, J.E. Richardson, and E.J. Shekita, “Object and file management in the EXODUS extensible database system,” Proc. 12th Int’l Conf. on Very Large Data Bases, pp. 91-100, August 1986.

[3] P. Cheeseman and J. Stutz, “Bayesian classification (AutoClass): theory and results,” in U.M. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pp. 153-180, AAAI/MIT Press, 1996.

[4] S.-C. Chen and R. L. Kashyap, “A Spatio-Temporal Semantic Model for Multimedia Presentations and Multimedia Database Systems,” accepted for publication in *IEEE Transactions on Knowledge and Data Engineering*, 2000.

[5] M. Ester, H.P. Kriegel, and X. Xu, “Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification,” Proc. Fourth Int’l Symp. Large Spatial Databases (SSD’95), pp. 67-82, August 1995.

[6] K.A. Hua, S.D. Lang, and W.K. Lee, “A decomposition-based simulated annealing technique for data clustering,” Proc. of the Thirteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), 1994.

[7] S.J. Kim, J. Baberjee, W. Kim, and J.F. Garza, “Clustering a DAG for CAD databases,” IEEE Transactions on Software Engineering, 14(11), pp. 1684-1699, November 1988.

[8] H. Lu, R. Setiono, and H. Liu, “NeuroRule: A connectionist approach to data mining,” Proc. 21th Int’l Conf. Very Large Data Bases, pp. 478-489, September 1995.

[9] K. Shannon and R. Snodgrass, “Semantic clustering,” in A. Dearle, G.M. Shaw, and S.B. Zdonik, editors, Implementing Persistent Object Bases: Principles and Practice, pp. 389-402, Morgan Kaufmann, 1991.

[10] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, “Database Clustering and Data Warehousing,” in *1998 ICS Workshop on Software Engineering and Database Systems*, pp. 30-37, Dec. 17-19, 1998.

[11] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, “Information Retrieval Using Markov Model Mediators in Multimedia Database Systems,” 1998 International Symposium on Multimedia Information Processing, pp. 237-242, Dec. 14-16, 1998.

[12] R. Srikant and R. Agrawal, “Mining generalized association rules,” Proc. 21th Int’l Conf. Very Large Data Bases, pp. 407-419, September 1995.

[13] R. Srikant and R. Agrawal, “Mining quantitative association rules in large relational tables,” Proc. 1996 ACM SIGMOD Int’l Conf. Management Data, pp. 1-12, June 1996.

[14] M.M. Tsangaris and J.F. Naughton, “On the performance of object clustering techniques,” Proc. ACM SIGMOD Int’l Conf. on Management of Data, pp. 144-153, June 1992.

[15] G. Wiederhold, “Mediators in the architecture of future information systems,” IEEE Computer, pp. 38-49, March 1992.

[16] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: An efficient data clustering method for very large databases,” Proc. 1996 ACM SIGMOD Int’l Conf. Management Data, June 1996.