Organizing a Network of Databases Using Probabilistic Reasoning

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

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

R.L. Kashayp
Purdue University
School of Electrical
and Computer Engineering
West Lafayette, IN 47907
kashyap@ecn.purdue.edu

Abstract

Due to the complexity of real-world applications, the number of databases and the volumes of data in databases have increased tremendously. With the explosive growth in the amount and complexity of data, how to effectively organize the databases and utilize the huge amount of data becomes important. For this purpose, a probabilistic network that organizes a network of databases and manages the data in the databases is proposed in this paper. Each database is represented as a node in the probabilistic network and the affinity relations of the databases are embedded in the proposed *Markov model mediator (MMM)* mechanism. Probabilistic reasoning technique is used to formulate and derive the probability distributions for an MMM. Once the probability distributions of each MMM are generated, a stochastic process is conducted to calculate the similarity measures for pairs of databases. The similarity measures are transformed into the branch probabilities of the probabilistic network. Then, the data in the database can be managed and utilized to allow user queries for database searching and information retrieval. An example is included to illustrate how to model each database into an MMM and how to organize the network of databases into a probabilistic network.

1 Introduction

The exponential growth of computer networks and data-collection technology has generated an incredibly large offer of products and services for the users of the computer networks. Such a large-scaled network of databases may consists of multiple, autonomous, and interconnected operational databases, that is, some are relational, some are object-oriented, some are hierarchical, and some are multimedia databases. In addition, the number of databases and the volumes of data in databases have increased tremendously in such an information-providing environment. Hence, there is the need for a *database management system (DBMS)* that has the capabilities to provide a suitable environment for storing, managing, and retrieving data in the database systems.

With the explosive growth in the amount and complexity of data, the need to extend the database technology to effectively manage the databases and utilize the large amount of data has posed a great challenge to the database research community. Toward this end, a probabilistic network based mechanism is proposed for managing a network of databases. Each node in the probabilistic network is an individual database. The mechanism is based upon a core set of database constructs and a set of queries with the probabilistic descriptions of database access patterns. The probabilistic descriptions of the queries are used to generate the training traces. The training traces and the semantic structures of the databases incorporate probabilistic reasoning techniques to construct the probabilistic network.

The network of databases is modeled as a probabilistic network with the affinity relations of the databases embedded in some probabilistic models. A mathematically sound framework, called the *Markov model mediator (MMM)* mechanism, serves as the probabilistic model for each node in the probabilistic network. MMMs adopt 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; while a mediator is a program that collects information from one or more sources, processes and combines it, and exports the resulting information [10] [11]. Many applications use Markov model as a framework such as *Hidden Markov Models (HMMs)* which

0-7803-6583-6/00/$10.00 © 2000 IEEE

are based on modeling patterns as a sequence of observation vectors derived from a probabilistic function of a non-deterministic first-order Markov process in speech recognition [6], and Markov Random Field Models permit the introduction of spatial context into pixel labeling problems and lead to algorithms for generating textured images, classifying textures, and segmenting textured images [4] [5] [7]. However, to our best knowledge, there is no existing research that uses Markov models as a framework in designing a database management system.

In our previous studies, the *MMM* mechanism has been used to facilitate the functionality of the database management systems [8] [9]. The semantic structure of each database is modeled by the sequence of the MMM states connected by transitions. Since an MMM possesses the stochastic property of the Markov models, the construction of the probabilistic network is based on complex statistical and probabilistic analyses which are best understood by examining the network-like structure in which those statistics are stored. With the help of probabilistic models, methods can be developed to manage the databases via probabilistic reasoning. For this purpose, a probabilistic reasoning approach based on the affinity measures of the databases is used to derived the three probability distributions for an MMM. Then, a stochastic process using these distributions is proposed to build the probabilistic network. The similarity values for the pairs of databases are calculated via the stochastic process and then transformed into the branch probabilities of the probabilistic network. A simple example is presented to show how the probabilistic reasoning technique is applied and how the stochastic process is executed for the construction of the probabilistic network. Since the MMMs possess the stochastic property of Markov models, database searching and information retrieval for queries can be processed via utilizing some designed stochastic processes. Therefore, the probabilistic network organizes the databases so that the data in the database can be managed and utilized to allow user queries for information retrieval.

The rest of the paper is organized as follows. The MMM mechanism is briefly introduced in Section 2. The construction of the probabilistic network is presented in Section 3. The probability reasoning technique, the stochastic process, and an example are also included in Section 3. The conclusions are drawn in Section 4.

2 Probabilistic Model – the Markov Model Mediator (MMM) Mechanism

Each individual database is modeled as an MMM that is represented by a 6-tuple $\lambda = (\mathcal{S}, \mathcal{F}, \mathcal{A}, \mathcal{B}, \Pi, \Psi)$.

1. $\mathcal{S}$ is a set of media objects called states.

   Each node in an MMM represents a *media object* that is used to denote the primitive constructed or manipulated entities in the multimedia databases [1]. Each individual database has its own set of media objects. An MMM consists of a sequence of states which represent the media objects in the databases. The states are connected by directed arcs (transitions) which contain probabilistic and other data used to determine which state should be selected next. Essentially, an MMM is a stochastic finite state machine with a stochastic output process attached to each state to describe the probability of occurrence of the output symbols (states).

2. $\mathcal{F}$ is a set of attributes/features.

   Since different media objects may have different types of attributes or features, each media object has its own set of attributes/features. $\mathcal{F}$ consists of all distinct attributes/features in the databases.

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

   $\mathcal{A}$ indicates the probabilities that go from one state (media object) to another.

4. $\mathcal{B}$ is the observation symbol probability distribution.

   $\mathcal{B}$ denotes the probability of observing an output symbol (attribute/feature) from a state (media object). Since a media object has its own set of attributes/features and an attribute (a feature) can belong to multiple media objects, the observation symbol probabilities show the probabilities an attribute (a feature) is observed from a set of media objects.

5. Π is the initial state probability distribution.

   Π gives the probability that a state (media object) can be the initial state for the incoming queries. Since the information from the training traces is available, the preference of the initial states for queries can be obtained.

6. Ψ is a set of multimedia augmented transition networks (ATNs).

   The multimedia ATN is a semantic model that is based on the ATN model. The multimedia ATN model has been proposed for multimedia presentations, multimedia database searching, and multimedia browsing [2] [3]. The arcs in an ATN represent the time flow from one state node to another. An arc represents an allowable transition from the node at its tail to the node at its head, and the labeled arc represents the transition function. An input string is accepted by an ATN if there is a path of transitions which corresponds to the sequence of symbols in the string and which leads from a specified initial state to one of a set of specified final states. In an MMM, each node (media object) is associated with an ATN.

3 Construction of Probabilistic Network

$\mathcal{A}, \mathcal{B}$, and Π are the probability distributions for an MMM. They play the major roles in the construction of the probabilistic network. The probabilistic reasoning technique is used to formulate and derive these three probability distributions for an MMM. The relative affinity values are used to construct the state transition probability distribution $\mathcal{A}$. The observation symbol probability distribution $\mathcal{B}$ captures the structure of the databases. Then, the training traces are used to construct the initial state probability distribution Π. The elements in $\mathcal{S}$ and $\mathcal{F}$ determine the dimensions of $\mathcal{A}$ and $\mathcal{B}$. Once $\mathcal{A}, \mathcal{B}$, and Π of each MMM are generated, a stochastic process is conducted to calculate the similarity measures for pairs of databases and to construct the probabilistic network.

3.1 Probabilistic Reasoning

To formulate the state transition probability distribution $\mathcal{A}$, the relative affinity values for pairs of media objects in a database are calculated. A relative affinity value between two media objects indicates how frequently these two media objects have been accessed together. If two media objects have a higher relative affinity relationship, the probability that a traversal choice to one state under a given state should be higher. The relative affinity value between two media objects m and n ($aff_{m,n}$) in a database d_i is defined as follows.

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

where q is the total number of queries, $access_k$ is the access frequency of query k, and $use_{k,m}$ is the usage pattern of a media object with respect to a query. $use_{k,m}$ (or $use_{k,n}$) has a value **1** if media object m (or n) is accessed by query k and has a value **0** otherwise. Based on the relative affinity values, the conditional probability ($a_{m,n}$) of traversing the state (media object) n given that the current state (media object) is m is referred as the state transition probability.

$$a_{m,n} = (f_{m,n}) \ / \ (f_m),$$

where $f_{m,n}$ denotes the joint probability defined as $(aff_{m,n}) \ / \ (\sum_m \sum_n aff_{m,n})$, and f_m is the marginal probability defined as $\sum_n f_{m,n}$.

Next, to formulate the observation symbol probability distribution $\mathcal{B}$, a temporary matrix BB is first generated using the following manner.

Let $m_1, m_2, \ldots, m_{n_i}$ be the media objects in d_i and $z_1, z_2, \ldots, z_{tot}$ be the attributes/features in all databases, where n_i is the number of media objects in d_i and tot is the number of distinct attributes/features in all databases. Put $m_1, m_2, \ldots, m_{n_i}$ in the columns of BB and $z_1, z_2, \ldots, z_{tot}$ in the rows of BB. Thus, BB is of size tot by n_i. Each entity of BB has a value **1** if the attribute/feature appears in the corresponding media object and has a value **0** otherwise. Then the observation symbol probability distribution $\mathcal{B}$ can be obtained via normalizing BB per column.

Finally, the initial state probability distribution Π_i for database d_i is defined as the fraction of the number of occurrences of media object m with respect to the total number of occurrences for all the member media objects in d_i from the training traces.

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

Again, q is the total number of queries, n_i is the number of media objects in d_i, and $use_{k,m}$ (or $use_{k,l}$) is the usage pattern of the media object m (or l) with respect to query k.

3.2 Stochastic Process

To construct the probabilistic network for a network of databases, first the similarity values for pairs of databases are calculated. The similarity value of two databases d_i and d_j is denoted by $SM(d_i, d_j)$ that indicates how well d_i and d_j match the observations generated by the sample queries.

Let $\mathcal{OS}$ be the set of all observations, O^k be an observation set with the attributes/features involved in query k, X be a set of media objects belonging to d_i in O^k, Y is a set of media objects belonging to d_j in O^k, $k1$ be the number of attributes/features belonging to the media objects in X, $k2$ be the number of attributes/features belonging to the media objects in Y, $N_k = k1 + k2$, and $F(N_k) = 10^{N_k}$ be an adjusting factor. Assume 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. Then $SM(d_i, d_j)$ is formulated as follows.

$$
\begin{aligned}
&SM(d_i, d_j) \\
&=(\textstyle\sum_{O^k \in \mathcal{OS}} P(O^k \mid X, Y; d_i, d_j) P(X, Y; d_i, d_j)) F(N_k) \\
&= (\textstyle\prod_{u=2}^{k1} \underbrace{P(x_u \mid x_{u-1})}_{A_i(x_u|x_{u-1})} \underbrace{P(x_1)}_{\Pi_i(x_1)}) \\
&\times (\textstyle\prod_{v=k1+2}^{N_k} \underbrace{P(y_{v-k1} \mid y_{v-k1-1})}_{A_j(y_{v-k1}|y_{v-k1-1})} \underbrace{P(y_1)}_{\Pi_j(y_1)}) \\
&\times (\textstyle\prod_{u=1}^{k1} \underbrace{P(o_u \mid x_u)}_{B_i(o_u|x_u)}) \\
&\times (\textstyle\prod_{v=k1+1}^{N_k} \underbrace{P(o_v \mid y_{v-k1})}_{B_j(o_v|y_{v-k1})}) \\
&\times (10^{N_k}).
\end{aligned}
$$

3.3 An Example

A simple example with four databases is used to illustrate how to organize these databases into a probabilistic network. Assume there are four databases d_1 to d_4, where d_1 has two media objects, d_2 has three media objects, d_3 has four media objects, and d_4 has four media objects.

d_1 = {*center*, *department*}
d_2 = {*dept*, *emp*, *proj*}
d_3 = {*employee*, *secretary*, *engineer*, *manager*}
d_4 = {*InletValve*, *NeedleSeat*, *InletNeedle*, *Maker*}

Let the media objects be numbered from *1* to *13* and the media objects in the same database have consecutive numbers. For example, the media objects *center* and *department* in d_1 have numbers *1* and *2*. The media objects *dept*, *emp*, and *proj* in d_2 have numbers *3*, *4*, and *5*, respectively. Similarly, the media objects in d_3 and d_4 can be numbered using the same manner.

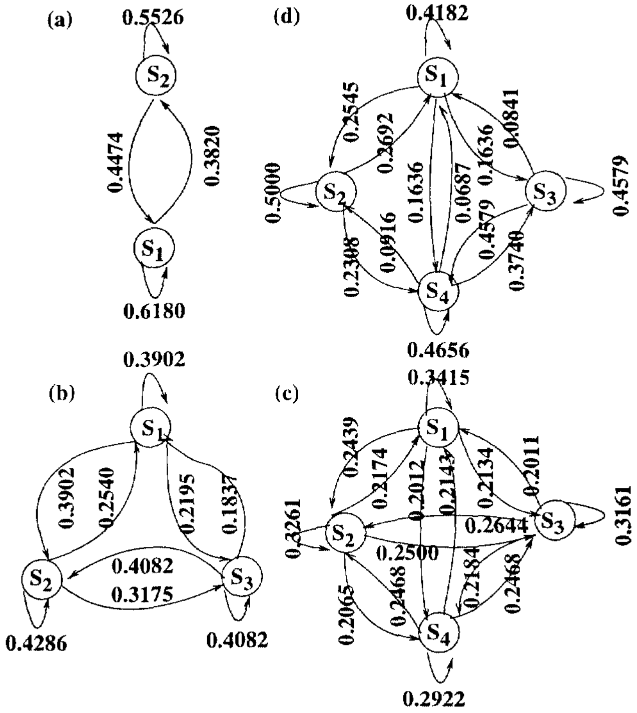

Figure 1: (a), (b), (c), and (d) are the constructed MMMs for the four databases d_1, d_2, d_3, and d_4, respectively.

A set of eight queries whose probabilistic descriptions including the database access patterns and the access frequencies of the queries are used to generate the training traces. Based on the information in the training traces, the three probability distributions for each database can be obtained via the proposed probabilistic reasoning approach. Figure 1 shows the constructed MMMs for the four databases. Figures 1(a), (b), (c), and (d) are the MMMs for d_1, d_2, d_3, and d_4, respectively. As can be seen from this figure, each database is modeled by an MMM and the state transition probabilities of each database are attached to the arcs of its MMM.

Based on the three probabilistic distributions for the MMMs and the set of sample queries, the similarity values for the pairs of databases are calculated. The resulting similarity values for the databases are shown in Table 1. As can be seen from Table 1, the similarity values are symmetric which means $SM(d_i, d_j)$ is equal to $SM(d_j, d_i)$. For example, the similarity value between d_1 and d_2 is *5.77* and the same value is shown between d_2 and d_1. The similarity values in

Table 1 are then transformed into the branch probability $P_{i,j}$ for nodes i and j (as shown in Table 2) in the probabilistic network. The transformation is executed by normalizing the similarity values per row to indicate the branch probabilities from a specific node (database) to all its accessible nodes (databases). In this example, the branch probabilities from d_1 to d_2, d_3, and d_4 are *0.1452*, *0.3223*, and *0.5325*, respectively.

Table 1: The similarity values for pairs of databases.

$SM(d_i, d_j)$	d_1	d_2	d_3	d_4
d_1	-	5.77	12.81	21.16
d_2	5.77	-	6.08	3.76
d_3	12.81	6.08	-	15.31
d_4	21.16	3.76	15.31	-

Table 2: The branch probabilities transformed from the similarity values (in Table 1).

$P_{i,j}$	d_1	d_2	d_3	d_4
d_1	-	0.1452	0.3223	0.5325
d_2	0.3696	-	0.3895	0.2409
d_3	0.3746	0.1778	-	0.4477
d_4	0.5260	0.0935	0.3806	-

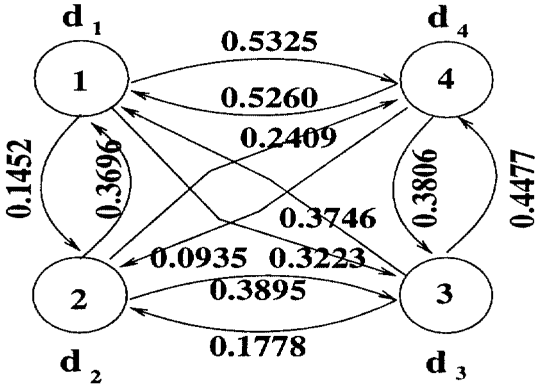

Figure 2: The probabilistic network for the four databases with nodes 1 to 4 representing databases d_1 to d_4, respectively.

From the obtained branch probabilities, the probabilistic network for the four databases can be constructed. Figure 2 gives the constructed probabilistic network with each node representing a database and each branch probability $P_{i,j}$ attached to the corresponding arc.

4 Conclusions

The emergence of networks of databases and the explosive growth in the sizes of networks and data have motivated the need for a good *database management system (DBMS)*. Toward this end, a probabilistic network approach that incorporates the probabilistic reasoning technique into a DBMS is proposed in this paper. The proposed probabilistic network-based approach manages and utilize the data in the databases in an information-providing database environment.

With the help of a probabilistic network, the affinity relation of the databases in the network can be embedded in the proposed MMM mechanism. The MMM mechanism serves as the probabilistic model for the nodes (databases) in the probabilistic network. To our knowledge, no existing research uses Markov models as a framework in designing a DBMS. Probabilistic reasoning technique is used to formulate and derive the probability distributions for the MMM mechanism. Probabilistic reasoning is powerful in a complex probabilistic network with a large number of states. Then, a stochastic process using these distributions is proposed to build the probabilistic network.

An example to illustrated how to model each database into an MMM and how to organize the network of databases into a probabilistic network is presented. Since the construction of the probabilistic network is based upon the core set of database constructs, the probabilistic descriptions of database access patterns, and the access frequencies of the queries, database searching and information retrieval for queries can be performed via some designed stochastic processes. In other words, the network of databases is organized into the probabilistic network in the way that the data in the database can be managed and utilized to allow user queries.

References

[1] K.S. Candan, P.V. Rangan, and V.S. Subrahmanian, "Collaborative Multimedia Systems: Synthesis of Media Objects," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 10, No. 3, pp. 433-457, May/June 1998.

[2] S-C. Chen and R. L. Kashyap, "A Spatio-Temporal Semantic Model for Multimedia Presentations and Multimedia Database Systems," accept for publication, *IEEE Transactions on Knowledge and Data Engineering*, 2000.

[3] S.-C. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "Augmented Transition Networks as Video Browsing Models for Multimedia Databases and Multimedia Information Systems," *the 11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'99)*, pp. 175-182, November 1999.

[4] P.J. Diggle, *Statistical Analysis of Spatial Point Patterns.* Academic Press, New York, 1983.

[5] O. Frank and D. Strauss, "Markov Graphs," *Journal of the American Statistical Association*, 81, 1986, pp. 832-842.

[6] L.R. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, 3(1), pp. 4-16, January 1986.

[7] B.D. Ripley, *Spatial Statistics.* John Wiley, Chichester, 1981.

[8] 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, December 1998.

[9] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "A Probabilistic-Based Mechanism For Video Database Management Systems," accept for publication, *IEEE International Conference on Multimedia and Expo (ICME 2000)*, July 30-August 2, 2000.

[10] G. Wiederhold, "Mediators in the architecture of future information systems," *IEEE Computer*, pp. 38-49, March 1992.

[11] G. Wiederhold, "Intelligent integration of information," in *ACM SIGMOD Conference*, pp. 434-437, May 1993.