JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 15, 217-241 (1999)

Design of a New Indexing Organization for A Class-Aggregation Hierarchy in Object-Oriented Databases[1]

CHIEN-I LEE, YE-IN CHANG* AND WEI-PANG YANG**
Institute of Information Education
National Tainan Teachers College
Tainan, Taiwan 700, R.O.C.
E-mail:leeci@ipx.ntntc.edu.tw
**Department of Applied Mathematics*
National Sun Yat-Sen University
Kaohsiung, Taiwan 804, R.O.C.
E-mail: changyi@math.nsysu.edu.tw
***Department of Computer and Information Science*
National Chiao Tung University
Hsinchu, Taiwan 300, R.O.C.
E-mail: wpyang@cis.nctu.edu.tw

In an object-oriented database, a class consists of a set of attributes, and the values of the attributes are objects that belong to other classes; that is, the definition of a class forms a class-aggregation hierarchy of classes. A branch of such a hierarchy is called a *path*. Several index organizations have been proposed to support object-oriented query languages, including *multiindex, join index, nested index* and *path index*. All the proposed index organizations are helpful only for a query which retrieves the objects of the root class of a given path using a predicate which specifies the value of the attribute at the end of the path. In this paper, we propose a new index organization for evaluating queries, called *full index*, where an index is allocated for each class and its attribute (or nested attribute) along the path. From the analysis results, we show that a *full index* can support any type of query along a given path with a lower retrieval cost than all the other index organizations. Moreover, to reduce the high update cost for a long given path, we split the path into several subpaths and allocate a separate index to each subpath. Given a path, the number of subpaths and the index organization of each subpath define an index configuration. Since a low retrieval cost and a low update cost are always a trade-off in index organizations, we also propose cost formulas to determine the index configuration which can provide the best performance for various applications by taking into account various types of queries along a given path and a set of queries with more than one nested predicate along a given path.

Keywords: access methods, complex objects, index selection, object-oriented databases, query optimization

Received April 10, 1997; accepted December 14, 1997.
Communicated by Arbee L. P. Chen.
[1]This research was supported in part by the National Science Council of Republic of China under Grant No. NSC-84-2213-E-110-009.

1. INTRODUCTION

The new generation of computer-based applications, such as computer-aided designed and manufacturing (CAD/CAM), multimedia databases (MMDB), and software development environments (SDEs), requires more powerful techniques to generate and manipulate large amounts of data. The traditional well-known record-based relational data model does not provide the possibility of directly modeling complex data. Moreover, many complex relationships among data, for example, instantiation, aggregation and generalization, can not be well defined in the relational data model. Furthermore, the relational data model does not provide mechanisms to associate data behavior with data definitions at the schema level.

Object-oriented database management systems [1, 10, 12, 13, 19, 20, 22-24] represent one of the most promising directions in the database area for meeting the requirements of advanced applications. An object-oriented data model not only provides great expressive power for describing data and defining complex relationships among data, but also provides mechanisms for behavioral abstraction. In an object-oriented data model [4, 6, 9], any real-world entity is represented by only one data modeling concept, the object. Each object is identified by a unique identifier (UID). The state of each object is defined at any point in time by the value of its attributes. The attributes can have as values both primitive objects (for example, strings, integers, or booleans) and non-primitive objects, which in turn, consist of a set of attributes. Objects with similar attributes are grouped into classes. A class C consists of a number of attributes, and the value of an attribute A of an object belonging to the class C is an object or a set of objects belonging to some other class C'. The class C' is called the *domain* of the attribute A of the class C, and this association is called an *aggregation relationship* between the classes C and C'. The class C' in turn consists of a number of attributes, whose domains are other classes. In general, a class is a hierarchy of classes of *aggregation relationships*, called an *aggregation hierarchy*. A branch of such a hierarchy is called a *path*. An example of a class-aggregation hierarchy is shown in Fig. 1. An example of a path against this class-aggregation hierarchy is *Student.study. taught-by.work-in.name*. Several index organizations [3, 5, 8, 14, 16-18, 21, 25] proposed to support object-oriented query languages, including *multiindex* [18], *join index* [21], *nested index* [3], *path index* [3] and *access support relation* [14].

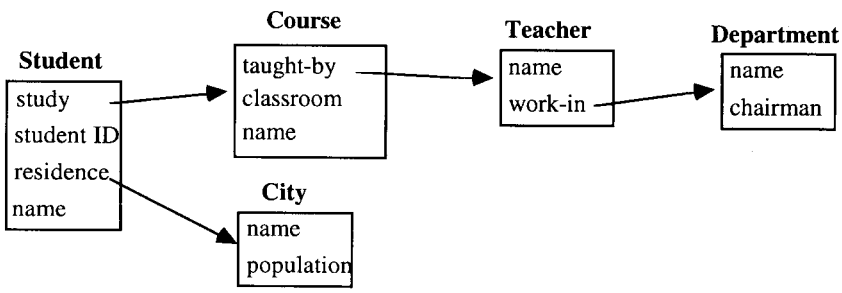

Fig. 1. An example of a class-aggregation hierarchy.

Consider the following query: *Retrieve all the students who study in some courses in the Department of Computer Science*, for the class-aggregation hierarchy shown in Fig. 1 and the related instances shown in Fig. 2, where 'study in the Department of Computer Science' is a nested predicate. (Note that nested predicates are often expressed using *path-expressions*; the above nested predicate can be expressed as *Student.study.taught-by.work-in.name* = 'Computer Science'.) Given a path = *Student.study.taught-by.work-in.name*, there are four indices in a *multiindex* set. The first index is on the subpath *Student.study* and contains the following pairs: (Course[i], {Student[o]}), (Course[j], {Student[p]}) and (Course[k], {Student[q]}). The second index is on the subpath *Course.taught-by* and contains the following pairs: (Teacher[i], {Course[k], Course[l]}) and (Teacher[j], {Course[i], Course[j]}). The third index is on the subpath *Teacher.work-in* and contains the following pairs: (Department[l], {Teacher[i]}) and (Department[m], {Teacher[j]}). The forth index is on the subpath *Department.name* and contains the following pairs: (Computer Science, {Department[l]}) and (Mathematics, {Department[m]}).

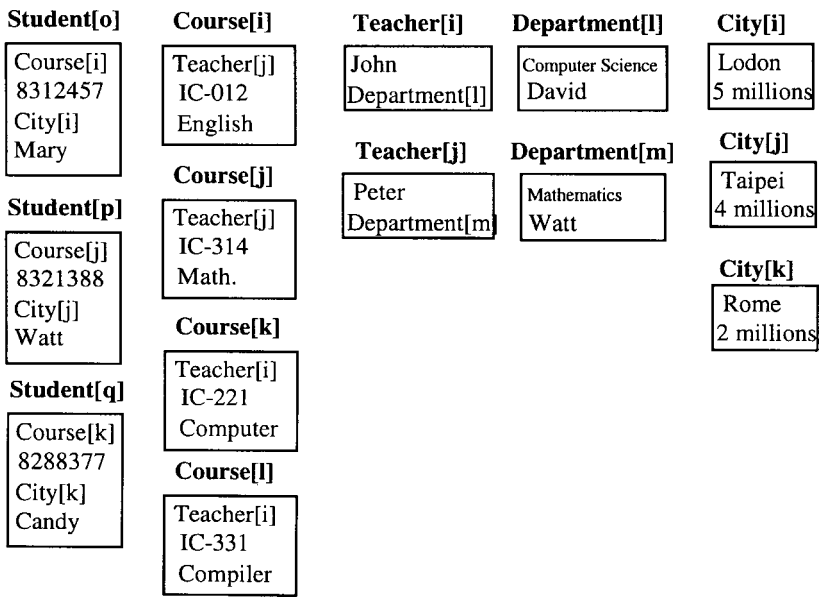

Fig. 2. Instances of classes in Fig. 1.

A *join index* [21] is similar to a *multiindex* except that a *join index* supports both forward and reverse traversal along the path; that is, two indices are allocated between each class and its immediate attribute along the path. That is, for a given path, the number of indices for a *join index* is two times that for a *multiindex*.

For the same example shown above, a *nested index* will contain the pairs (Computer Science, {Student[q]}) and (Mathematics, {Student[o], Student[p]}) while a *path index* will contain the following pairs:

(Computer Science, {Student[q].Course[k].Teacher[i].Department[l]}) and
(Mathematics, {Student[o].Course[i].Teacher[j].Department[m],
Student[p].Course[j].Teacher[j].Department[m]}).

An *access support relation* [14] is an organization very similar to a path index except that an *access support relation* can store incomplete path instantiations using null values in relations.

In general, a *multiindex* [18] is allocated on each class traversed by the path, which solves a nested predicate by scanning a number of indices equal to the path-length. Therefore, a *multiindex* has a high retrieval cost but a low update cost. Since a *nested index* only associates instances of the first class of the path with the values at the end of the path, a *nested index* has a lower retrieval cost for querying on the first class with the nested predicate on the last attribute of the path but has a high update cost for forward and backward object traversals to access the database itself. A *path index* [3] provides an association between an object at the end of the path and the instantiations ending with the object. Therefore, *a path index* can be used to evaluate nested predicates on all classes along the path; however, a *path index* has a high update cost.

All the above proposed index organizations will be helpful only to a query which retrieves the objects of the root class of a given path using a predicate which specifies the value of the attribute at the end of the path. Consider another query: *Retrieve all the courses taught by those teachers who are in the Department[l]*. The path-expression for this query is *Course.taught-by.work-in* = 'Department[l]'. To answer this question, we have to scan the second index and the third index in the *multiindex* described above. The *nested index* described above will not be helpful in answering the query while the *path index* described above will be only able to answer with a partial result ({Course[k]}) by scanning all the *path index*. (Note that the answer to the query should be {Course[k], Course[l]}.) To reduce the high retrieval cost in a *multiindex* and to overcome the problem where some queries in a *nested index* and a *path index* cannot be answered, in this paper, we propose a new index organization for evaluating queries, called a *full index*, where an index is allocated for each class and each of its immediate and nested attributes along the path. For the same example shown above, a *full index* will contain 10 indices as shown in Table 1.

Table 1. An example of a full index.

Class	Attribute	Contents
Student	Study	(Course[i], {Student[o]}), (Course[j], {Student[p]}) and (Course[k], {Student[q]})
Student	Taught-by	(Teacher [i], {Student[q]}) and (Teacher[j], {Student[o], Student[p]})
Student	Work-in	(Department[l], Student[q]}) and (Department[m], {Student[o], Student[p]})
Student	Name	(Computer Science, {Student[q]}) and (Mathematics, {Student [o], Student[p]})
Course	Taught-by	(Teacher[i], {Course[k], Course[l]}) and (Teacher[j], {Course[i], Course[j]})
Course	Work-in	(Department[l], {Course[k], Course[l]})and (Department[m], {Course[i], Course[j]})
Course	Name	(Computer Science, {Course[k], Course[l]}) and (Mathematics, {Course[i], Course[j]})
Teacher	Work-in	(Department[l], {Teacher[i]}) and (Department[m], {Teacher[j]})
Teacher	Name	(Computer Science, {Teacher[i]}) and (Mathematics, {Teacher[j]})
Department	Name	(Computer Science, {Department[l]}) and (Mathematics, {Department[m]})

A comparison of a *multiindex*, a *nested index*, a *path index* and a *full index* is shown in Fig. 3. From the analysis results, we can show that a *full index* can support any type of query along a given path against a class-aggregation hierarchy with a lower retrieval cost than all the other index organizations. Therefore, our full index is suitable for queries against a given path, where queries on subpaths might not be *predictable*. To reduce the high update cost for a long given path, we can split the path into several subpaths and allocate a separate index to each subpath [7, 14, 15]. Given a path, the number of subpaths and the index organization of each subpath define an index configuration. But the increase of the number of indices for subpaths will also increase the retrieval cost for scanning a number of indices, which results in a high retrieval cost. Since a low retrieval cost and a low update cost are always a trade-off in index organizations, we can establish a cost formula to look for a compromise between these two requirements. In [7], the authors proposed cost formulas to evaluate the costs of various index configurations. However, they do not support *partial instantiations* (defined in Section 2) and do not consider more than one nested predicate along the path. In this paper, we also propose cost formulas to determine the index configuration which can provide the best performance for various applications by taking into account various types of queries along a given path and a set of queries with more than one nested predicate along a given path.

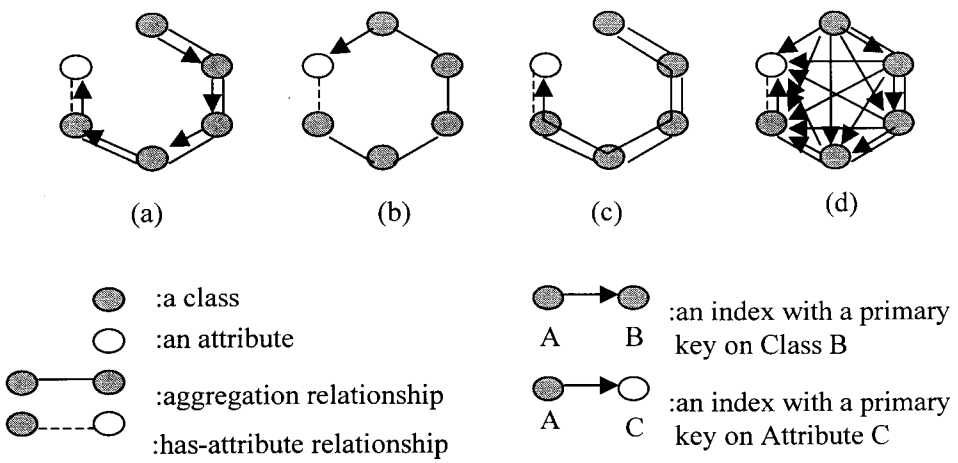

Fig. 3. A comparison: (a) a *multiindex*; (b) a *nested index*; (c) a *path index*; (d) a *full index*.

The rest of this paper is organized as follows. In Section 2, we define different types of queries along a given path. In Section 3, we introduce the proposed *full index* and related index operations. In Section 4, we present the cost model for a *full index* and give some analysis results of a *full index* compared with those of a *multiindex*, a *path index* and a *nested index*. In Section 5, we present cost formulas which determine an optimal index configuration. Finally, Section 6 concludes this paper.

2. QUERY TYPES IN A CLASS-AGGREGATION HIERARCHY

An attribute of any class on a class-aggregation hierarchy is logically an attribute of the root of the hierarchy; that is, the attribute is a nested attribute of the root class. A predicate on a nested attribute is called a *nested predicate*. A *path* is defined a $C(1).A(1).A(2).....A(n)$, where $C(i)$ is a class in a class-aggregation hierarchy and $A(i)$ is an attribute of class $C(i)$, $1 \le i \le n$. The *path-length* indicates the number of classes along a path; that is, the value is n. An *instantiation* of a path is defined as a sequence of $(n + 1)$ objects as $O(1).O(2).....O(n + 1)$, where $O(i)$ is an instance of class $C(i)$, $1 \le i \le (n+1)$. A *patial instantiation* of a path is defined as a sequence of objects as $O(i).O(i +1).....O(j)$, where $O(i)$ is an instance of class $C(i)$, $1 \le i \le j \le (n+1)$. Object-oriented query languages allow objects to be restricted by predicates on both nested and non-nested attributes of objects. In the following, we define types of queries along a given path against a class-aggregation hierarchy.

Definition 1: *Given a path which is defined as* $C(1).A(1).A(2)...A(n)$ $(n \ge 1)$ *and an aggregation hierarchy H, a simple type of query is expressed using path-expression as* $C(i).A(i).....A(j)$ = *'*$O(j)$*', where* $1 \le i \le j \le n$.

That is, a simple type of query retrieves the objects of a class along a given path using a predicate on its nested (or non-nested) attribute, where the specified class need not be the root of the path and the specified attribute need not be the end of the path. The following query **Q1** shows an example of a simple type of query against the class-hierarchy shown in Fig. 1.

Q1: *Retrieve all the students who study in the Department of Computer Science.*

Q1 contains the nested predicate 'study in the Department of Computer Science'. Nested predicates are often expressed using *path-expressions*. For example, the above nested predicate can be expressed as *Student.study.taught-by.work-in.name* = 'Computer Science'.

Definition 2: *Given a path which is defined as* $C(1).A(1).A(2).....A(n)$ $(n \ge 1)$ *and an aggregation hierarchy H, a k-degree complex type of query is expressed using path-expressions as*$C(i).A(i)....A(j_1)=$*'*$O(j_1)$*',*$C(i).A(i)....A(j_2)=$*'*$O(j_2)$*',....,*$C(i).A(i).....A(j_k)=$*'*$O(j_k)$*', where* $1 \le m \le k$, $1 \le i \le j_m \le n$.

That is, a *k*-degree complex type of query retrieves the objects of a class along a given path using *k* predicates on its *k* nested (or non-nested) attributes along the given path. The following query **Q2** shows an example of a 2-degree complex type of query against the class-hierarchy shown in Fig. 1, where the path-expressions are *Student.study.taught-by* = 'Teacher[i]' and *Student.study* = 'Course[i]'.

Q2: *Retrieve all the students who are taught by Teacher[i] and who study in Course[i].*

Definition 3: *Given a path which is defined as* $C(1).A(1).A(2).....A(n)(n \geq 1)$ *and an aggregation hierarchy H, a general set of queries is expressed using path-expressions as* $C(i_1).A(i_1)....A(j_1) = 'O(j_1)', C(i_2).A(i_2)....A(j_2) = 'O(j_2)', ..., C(i_k).A(i_k)....A(j_k) = 'O(j_k)'$, *where* $1 \leq m \leq k$, $1 \leq i_m \leq j_m \leq n$.

That is, there is more than one query along a given path. The following example **Q3** shows two queries in a general set against the class-hierarchy shown in Fig. 1, where their path-expressions are *Student.study* = 'Course[i]' and *Course.taught-by.work-in.name* = 'Mathematics', respectively.

Q3: *Retrieve all the students who study in Course[i], and retrieves all the courses taught by those teachers who are in the Department of Mathematics.*

3. A FULL INDEX

In this section, we first give the formal definition of a *full index*. Then, we describe four operations on a *full index*, which are *retrieval*, *update*, *insertion* and *deletion*.

3.1 Organization

Definition 4: *Given a path P which is defined as* $C(1).A(1).A(2)....A(n)(n \geq 1)$ *and an aggregation hierarchy H, a full index* (*FX*) *on P is defined as a set of indices, which are* $FX_1^1, FX_1^2, ..., FX_1^n, FX_2^2, ..., FX_n^n$, *where* FX_i^j *is an index on class* $C(i)$ *and attribute* $A(j)$, $1 \leq i \leq j \leq n$.

For example, let a path = *Student.study.taught-by.work-in.name* of *H* be that shown in Fig. 1 and the instances of classes in *H* be those shown in Fig. 2; a *full index* consisting of ten indices is described as follows.

The first index FX_1^1 on class *Student* and attribute *study* contains the following pairs:
(Course[i], {Student[o]}),
(Course[j], {Student[p]}) and
(Course[k], {Student[q]}).

The second index FX_1^2 on class *Student* and attribute *taught-by* contains the following pairs:
(Teacher[i], {Student[q]}) and
(Teacher[j], {Student[o], Student[p]}).

The third index FX_1^3 on class *Student* and attribute *work-in* contains the following pairs:
(Department[l], {Student[q]}) and
(Department[m], {Student[o], Student[p]}).

The fourth index FX_1^4 on class *Student* and attribute *name* of class *Department* contains the following pairs:
(Computer Science, {Student[q]}) and
(Mathematics, {Student[o], Student[p]}).

The fifth index FX_2^2 on class *Course* and attribute *taught-by* contains the following pairs:

(Teacher[i], {Course[k], Course[l]}) and
(Teacher[j], {Course[i], Course[j]}).
The sixth index FX_2^3 on class *Course* and attribute *work-in* contains the following pairs:
(Department[l], {Course[k], Course[l]}) and
(Department[m], {Course[i], Course[j]}).
The seventh index FX_2^4 on class *Course* and attribute *name* of class *Department* contains the following pairs:
(Computer Science, {Course[k], Course[l]}) and
(Mathematics, {Course[i], Course[j]}).
The eighth index FX_3^3 on class *Teacher* and attribute *work-in* contains the following pairs:
(Department[l], {Teacher[i]}) and
(Department[m], {Teacher[j]}).
The ninth index FX_3^4 on class *Teacher* and attribute *name* of class *Department* contains the following pairs:
(Computer Science, {Teacher[i]}) and
(Mathematics, {Teacher[j]}).
The tenth index FX_4^4 on class *Department* and attribute *name* contains the following pairs:
(Computer Science, {Department[l]}) and
(Mathematics, {Department[m]}).

3.2 Operations

A *full index* supports fast retrieval of all types of queries defined in Section 2. An evaluation of a nested predicate against the nested attribute $A(j)$ of class $C(i)$ requires a lookup of a single index FX_i^j, where $1 \le i \le j \le n$. Suppose that an instance $O(i)$ of class $C(i)$ along the path has an object $O(i+1)$ as the value of the attribute $A(i)$. Now, $O(i)$ is updated to a new object $O'(i+1)$. An update to the *full index* proceeds as follows.

First, we perform a lookup of index FX_i^i and replace $O(i+1)$ of $O(i)$ with a new object $O'(i+1)$. Second, we update the index FX_m^k using the index FX_m^{k-1} and FX_k^k, where $1 \le m < i$ and $i \le k \le n$. Third, we update the index FX_i^m using the index FX_i^{m-1} and FX_m^m, where $i < m \le n$.

As in the examples shown in Figs. 1 and 2, suppose the object in attribute *taught-by* of *Course*[*i*] is updated from *Teacher*[*j*] to *Teacher*[*i*]. Then, a series of update operations on the *full index* are performed as follows.

First, we perform a lookup of the index FX_2^2 and replace *Teacher*[*j*] of *Course*[*i*] with *Teacher*[*i*]; that is, FX_2^2 contains the following pairs:
(Teacher[i], {Course[i], Course[k], Course[l]}) and
(Teacher[j], {Course[j]}).
Second, we update the index FX_1^2 using FX_1^1 and FX_2^2; that is, FX_1^2 contains the following pairs:
(Teacher[i], {Student[o], Student[q]}) and
(Teacher[j], {Student[p]}).
Then, we update the index FX_1^3 using FX_1^2 and FX_3^3; that is, FX_1^3 contains the following pairs:

(Department[l], {Student[o], Student[q]}) and

(Department[m], {Student[p]}).

Moreover, we update the index FX_1^4 using FX_1^3 and FX_4^4; that is, FX_1^4 contains the following pairs:

(Computer Science, {Student[o], Student[q]}) and

(Mathematics, {Student[p]}).

Third, we update the index FX_2^3 using FX_2^2 and FX_3^3; that is, FX_2^3 contains the following pairs:

(Department[l], {Course[i], Course[k], Course[l]}) and

(Department[m], {Course[j]}).

Then, we update the index FX_2^4 using FX_2^3 and FX_4^4; that is, FX_2^4 contains the following pairs:

(Computer Science, {Course[i], Course[k], Course[l]}) and

(Mathematics, {Course[j]}).

Insertion and deletion operations are similar to update operations. We perform an insertion/deletion operation of an object on index FX_i^j instead of the replacement operation in the first step of the update to the index.

4. PERFORMANCE ANALYSIS

In this section, we will describe the cost model and analyze some performance results of a *full index*. Moreover, a comparison of the performance of these related indexing schemes will also be presented.

4.1 Cost Model

In this paper, we use a cost model which is similar to the one proposed in [3, 7], in which the data structure used to model indices is based on a *B-tree* [2, 11]. Similar to their model [3, 7], we assume that the values of attributes are uniformly distributed among the instances of the class, and that all the key values have the same length. However, in [3, 7], since a *nested index* and a *path index* can not support any query for partial instantiations, they need to make one more assumption: no partial instantiations; that is, each instance of a class $C(i)$ is referenced by instances of class $C(i - 1)$, $1 < i \leq n$. In this paper, we remove this assumption so as to support a query of any type for partial instantiations in a *full index* by taking into account 'NULL' values of instances. Given a path $C(1).A(1).....A(n)$, the parameters that we consider in the cost model are grouped as follows (table on next page).

Moreover, to compare these indexing schemes on the same basis, we use the same assumptions and parameters as those in [3] except that we consider the case where an attribute $A(i)$ has a set of values, instead of a single value, and the average number of values in a set is $m(i)$. Therefore, we use $DV(i)$ to denote the number of distinct values for $A(i)$ and $D(i) = \binom{DV(i)}{m(i)}$ to denote the number of distinct sets. In this case, when $A(i)$ has a single value (i.e., $m(i) = 1$), $DV(i) = D(i)$. That is, we consider a more general case in $A(i)$. It is straightforward to extend the cost models [3] of a multiindex, path index, and nested index to consider the case where an attribute has a set of values instead of a single value. In this paper, we do have use the extended cost models of these indexing schemes in our performance comparison described in section 4.5. Note that, here, to simplify our presentation of the performance analysis, we omit some parameters used in [3] since they can be

easily derived from the other parameters. The values of those parameters which are not critical to the comparison are here kept constant; these parameters were also kept constant in [3].

Parameter	Description
$DV(i)$	Number of distinct values held in attributes $A(i)$, including the NULL value, $1 \le i \le n$
$m(i)$	Average number of values for a set for attributes $A(i)$, $1 \le i \le n$.
$D(i)$	Number of distinct sets for attribute $A(i)$, $1 \le i \le n$; that is, $D(i) = \binom{DV(i)}{m(i)}$
$N(i)$	Cardinality of class $C(i)$, including the NULL value, $1 \le i \le n$.
$K(i)$	Average number of instances of class $C(i)$ with the same set of values for attribute $A(i)$; that is, $K(i) = \left\lceil \frac{N(i)}{D(i)} \right\rceil$
$UIDL$	Length of the object-identifier in bytes.
PS	Page size in bytes.
d	Order of a nonleaf node.
f	Average fanout from a nonleaf node.
pp	Length of page pointer.
kl	Average length of a key value for the indexed attribute.
ol	Length of a header in an index record.
DS	Length of the directory at the beginning of the record, when the record size is greater than the page size.

4.2 Retrieval Cost

Let $K(i, j)$ be the average number of instances of class $C(i)$ having the same set of values held in the nested attribute $A(j)$, where $1 \le i \le j \le n$; that is, $K(i,j) = \prod_{r=i}^{j} K(r)$. XF_i^j denotes the average length of a leaf-node index record for the index FX_i^j in the *full index* and

$$XF_i^j = K(i,j)m(i)UIDL + kl + ol, \qquad XF_i^j \le PS,$$
$$XF_i^j = K(i,j)m(i)UIDL + kl + ol + DS, \quad XF_i^j > PS,$$
$$\text{where } DS = \left\lceil \frac{K(i,j)m(i)UIDL + kl + ol}{PS} \right\rceil (UIDL + pp).$$

The number of leaf pages LP_i^j for the index FX_i^j is

$$LP_i^j = \left\lceil \frac{D(j)}{\left\lfloor \frac{PS}{XF_i^j} \right\rfloor} \right\rceil, \qquad XF_i^j \le PS,$$

$$LP_i^j = D(j) \left\lceil \frac{XF_i^j}{PS} \right\rceil, \qquad XF_i^j > PS.$$

The number of index pages RC_i^j accessed in the index FX_i^j for a nested predicate on class $C(i)$ with a nested attribute $A(j)$ is

$$RC_i^j = h(i, j) + 1, \qquad XF_i^j \le PS,$$

where $h(i, j) (= \lceil \log_f D(j) \rceil)$ is the number of nonleaf nodes that must be accessed in the index FX_i^j. When the record size is larger than the page size, i.e., $XF_i^j > PS$, np is the number of leaf pages needed to store the record, i.e., $np = \left\lceil \frac{XF_i^j}{PS} \right\rceil$. Therefore,

$$RC_i^j = h(i, j) + np, \qquad XF_i^j > PS.$$

4.3 Maintenance Cost

The index maintenance cost derived from update, deletion or insertion operations for an instance of a class $C(i)$ is denoted by U, D and I, respectively. To simplify the analysis, we consider only the costs of leaf-page modifications and exclude the costs of index page splits. The cost CBM_i^i of an update on the index FX_i^i is the sum of the cost of removing the *UID* of object $O(i)$ from the record associated with its attribute $O(i + 1)$ and the cost of adding it to the new value $O'(i + 1)$; that is,

$$CBM_i^i = CO\ (1+pl),$$

where CO denotes the cost of finding the leaf node containing the key value and the cost of reading and writing the leaf node, and pl is the probability that the old and new values are on different leaf nodes.

When a leaf page is modified, one page access is needed to read the leaf page containing the update record, and another page access is needed to write this page; in addition, $h(i, i)$ pages are accessed to determine the leaf node containing the record to be updated. Therefore,

$$CO = h(i, i) + 2, \qquad XF_i^i \le PS.$$

When the record size is larger than the page size and np is the number of leaf pages needed to store the record, i.e., $np = \left\lceil \frac{XF_i^j}{PS} \right\rceil$, there are $h(i, i) + 1$ pages which must be accessed to find the header of the record with the old value. From the header of the record, it is possible to determine the page from which a *UID* must be deleted or to which a *UID* must be added. If this page is different from the page containing the header of the record, a further page access must be performed. This probability is given by $\frac{np-1}{np}$. Therefore,

$$CO = h(i,i) + 2 + \frac{np-1}{np}, \qquad XF_i^i > PS.$$

The probability that the current and new values are on different leaf nodes is

$$pl = 1, \qquad XF_i^i > PS,$$
$$pl = 1 - \frac{\left\lfloor \frac{PS}{XF_i^i} \right\rfloor - 1}{D(i)-1}, \qquad XF_i^i \le PS.$$

Moreover, an update operation on the index FX_i^i will cause other associated indices to be updated as stated in Subsection 3.2. When an index FX_m^k is updated ($1 \le m \le i$, $i < k \le n$), the total cost for this update consists of the update cost CBM_m^k on the index FX_m^k and the retrieval cost 2 RC_k^k. (Note that one cost RC_k^k is for finding $O(k + 1)$ to determine which object $O(k + 1)$ for object $O(k)$ is to be updated, and the other cost RC_k^k is for finding $O'(k + 1)$ to determine which new object $O'(k + 1)$ is to be updated.)

Therefore, the total update cost U_i is

$$U_i = CBM_i^i + \sum_{m=1}^{i-1}\sum_{k=i}^{n}(CBM_m^k + 2RC_k^k) + \sum_{m=i+1}^{n}(CBM_i^m + 2RC_m^m).$$

Since there is only a deletion of an old value or an insertion of a new value on the index FX_i^i, pl is 0. Therefore, the cost of deletion D_i and the cost of insertion I_i are given by

$$D_i = I_i = CO + \sum_{m=1}^{i-1}\sum_{k=i}^{n}(CBM_m^k + 2RC_k^k) + \sum_{m=i+1}^{n}(CBM_i^m + 2RC_m^m).$$

4.4 Storage Cost

The number of nonleaf pages NLP_i^j for the index FX_i^j is

$$NLP_i^j = \left\lceil \frac{LO}{f} \right\rceil + \left\lceil \frac{\left\lceil \frac{LO}{f} \right\rceil}{f} \right\rceil + \ldots\ldots + \lceil X \rceil,$$

where $LO = \min(D(j), LP_i^j)$ and each term is successively divided by f until the last term X is less than f.

Then, the total storage cost SC for a *full index* is

$$SC = \sum_{i=1}^{n}\sum_{j=i}^{n}(LP_i^j + NLP_i^j).$$

4.5 A Comparison

In this subsection, we will show a number of interesting results of a *full index* (denoted as *FX*) on the basis of the analysis cost model described in the above subsections and compare the performance of the *full index* with that of a *multiindex* (denoted as *MX*), a *nested index* (denoted as *NX*) and a *path index* (denoted as *PX*) [3]. (Note that since a *join index* is

similar to a *multiindex* and an *access support relation* is similar to a *path index*, we omit the performance for a *join index* and that for an *access support relation*.) By using different values of parameters, we can simulate some interesting situations for different application requirements. However, some parameters are kept constant in all the simulations, namely, $N(1)$=200,000, $UIDL$=8, kl=2, ol=6, pp=4, f=218, d=146 and PS=4096. The values of these parameters are the same as those in [3].

Fig. 4 shows the retrieval costs for simple queries with a path-expression $C(1).A(1)....A(3)$ = '$O(3)$', using a *multiindex*, a *nested index*, a *path index* and a *full index*, respectively, where n = 3, $m(1) = m(2) = m(3) = 1$, $K(2) = K(3) = 10$, and $K(1)$ is varied from 1 to 50. Fig. 5 shows the retrieval costs for simple queries with a path-expression $C(1).A(1)....A(3)$ = '$O(3)$', using a *multiindex*, a *nested index*, a *path index* and a *full index*, respectively, where n = 3, $K(1) = K(2) = K(3) = 10$, $m(2) = m(3) = 1$, and $m(1)$ is varied from 1 to 5. From these two figures, we observe that the retrieval costs for these four indexes are increased with the values of $K(1)$ or $m(1)$. The reason is that as $K(3)$ or $m(3)$ increase, the size of a leaf-node index record increases, which may result in an increase of the number of leaf pages for the index record. Consequently, the retrieval costs increase. Moreover, since the *multiindex* requires scanning of three indices to access the desired objects, the *multiindex* has the highest retrieval cost. The *full index* has a lower retrieval cost than the *multiindex* and the *path index*. The reason is that the *full index* requires only one lookup in the index FX_1^3, but the *multiindex* requires one lookup for each index and the *path index* has to perform a lookup for an index larger than the *full index*. In this case, the *nested index* is the same as the index FX_1^3 of the *full index*. Therefore, the *nested index* has the same retrieval cost as the *full index*.

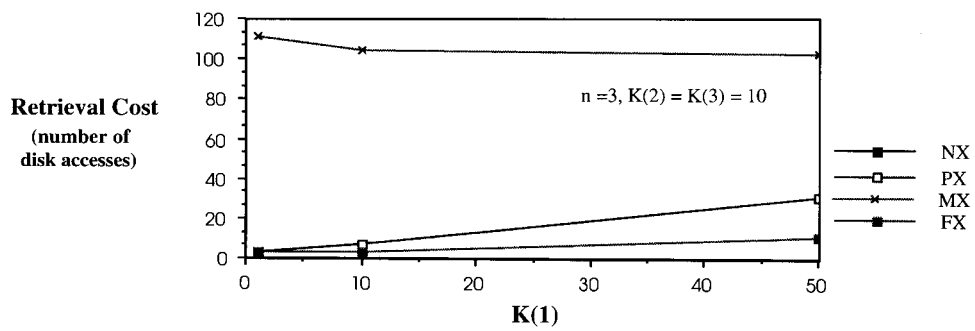

Fig. 4. The retrieval cost under different values of $K(1)$.

Fig. 6 shows the retrieval costs for simple queries with a path-expression $C(1).A(1)....A(n)$ = '$O(n)$', using a *full index*, a *multiindex*, a *nested index* and a *path index*, respectively, where $m(y) = 1$, $K(y) = 2$, $1 \le y \le n$ and n is varied from 2 to 10. For the same reasons in Fig. 4, the *multiindex* has the highest retrieval cost, and the *full index* and the *nested index* have the lowest retrieval cost. Moreover, since a *multiindex* is allocated to each class traversed by the path, which involves solving a nested predicate by scanning a number of indices equal to the path-length, the retrieval cost of the *multiindex* increases as n increases. A *path index* provides an association between an object at the end of the path and the instantiations

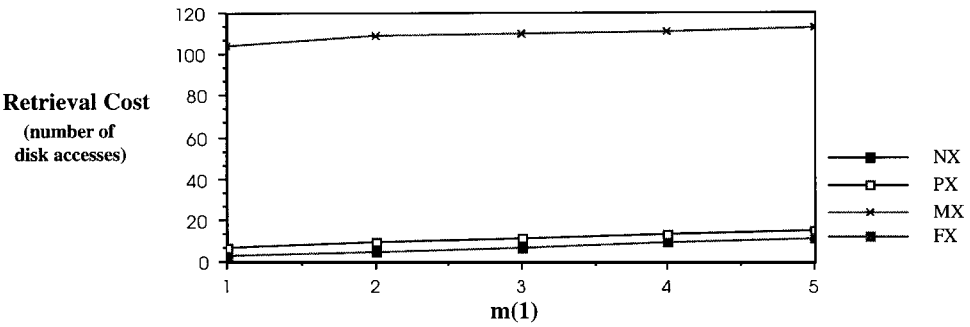

Fig. 5. The retrieval cost under different values of $m(1)$.

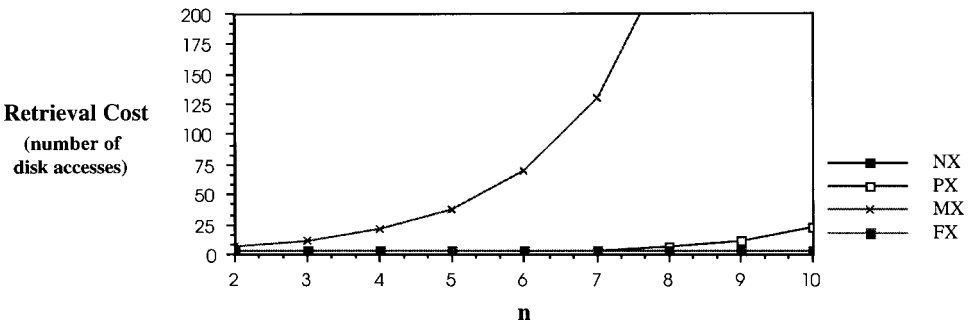

Fig. 6. The retrieval cost under different values of the path-length (n).

ending with the object. Therefore, the index size of a *path index* increases as n increases, which results in an increase of the retrieval cost. Since the index FX_1^n of a *full index* and a *nested index* only associates instances of the first class of the path with the values at the end of the path, a *full index* and a *nested index* have lower retrieval costs, and these costs are not affected by n.

Consider simple queries with path-expressions $C(1).A(1)....A(j) = 'O(j)'$, where $1 \le j \le n$, $m(y) = 1$, $K(y) = 2$, $1 \le y \le n$ and n is varied from 2 to 10. Suppose a *multiindex*, a *nested index*, a *path index* and a *full index* have been allocated for simple queries with a path-expression $C(1).A(1)....A(n) = 'O(n)'$, respectively. Since a *path index* and a *nested index* can not support any query for partial instantiations as stated before, a lookup in a real database is required. Therefore, the retrieval costs for a *nested index* and a *path index* will be very high. In Fig. 7, we compare the average retrieval cost based on a *multiindex* and a *full index* for all the queries with path-expressions $C(1).A(1)....A(j) = 'O(j)'$, where $1 \le j \le n$, which have the same probability. From this figure, we can find that a *full index* can provide a constant retrieval cost because only one lookup on an index is required for each query while the retrieval cost for a *multiindex* increases as n increases for because scanning of a number of indices is required for each query.

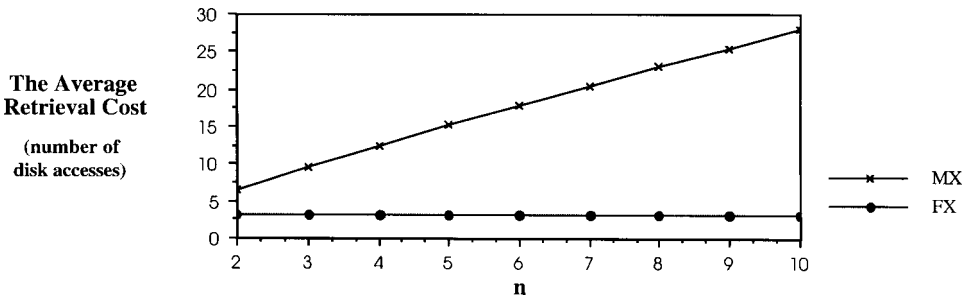

Fig. 7. The average retrieval cost for simple queries.

Fig. 8 shows the average retrieval cost for queries in a general set with path-expressions $C(i).A(i)....A(j) = 'O(j)'$, where $1 \le i \le n$, $i \le j \le n$ and n is varied from 2 to 10. For the same reasons in Fig. 7, we can find that a *full index* can provide a constant retrieval cost while the retrieval cost for a *multiindex* increases as n increases.

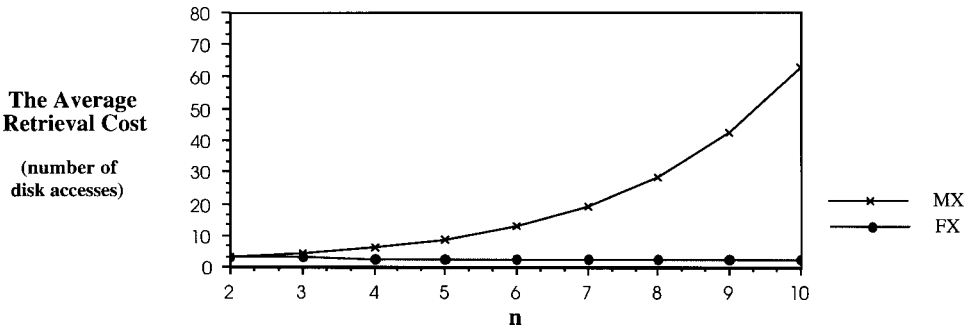

Fig. 8. The average retrieval cost for queries in a general set.

Fig. 9 shows the average update costs for a path = $C(1).A(1)....A(n)$, using a *full index*, a *multiindex*, a *nested index* and a *path index*, respectively, where $m(y) = 1$, $K(y) = 4$, $1 \le y \le n$ and n is varied from 2 to 10. A *multiindex* has the lowest average update cost because a single index is updated for each update operation. As stated in [3], for an update on an object $O(i)$, a *path index* requires the cost of a *forward traversal* and the cost of a *B-tree* update, and a *nested index* requires the cost of a *forward* and a *backward traversal* and the cost of a *B-tree* update. (Note that a *forward traversal* is defined as the accesses of objects $O(i + 1)$,..., $O(n)$ such that $O(i + 1)$ is referenced by object $O(i)$ through attribute $A(i)$, ..., and $O(n)$ is referenced by object $O(n–1)$ through attribute $A(n –1)$. On the other hand, a *backward traversal* is defined as the accesses of objects $O(i - 1)$,..., $O(2)$ such that $O(i –1)$ is referenced by object $O(i)$ through attribute $A(i –1)$, ..., and $O(1)$ is referenced by object $O(2)$ through attribute $A(1)$.) The cost for a *forward traversal* is in proportion to n, and the cost for a *backward traversal* is in proportion to 2^n[3]. Therefore, a *nested index* has a higher

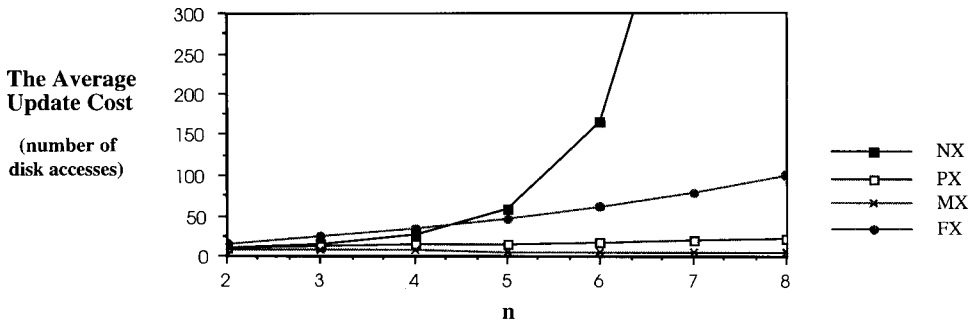

Fig. 9. The average update cost.

average update cost than a *path index*, and a *path index* has a higher average update cost than a *multiindex*. (Note that as stated in [3], the update cost of the traversal operations for a nested index is also related to the size of the real database involved in the path while the update cost of our full index is independent of the size of the real database involved in the path.) Since in a *full index*, an update operation in an index may cause some other update operations in other indices, the average update cost for a *full index* is higher than one for a *multiindex*. Moreover, the number of updated indices for an update in a *full index* is in proportion to n^2, and the cost for these updated indices is higher than the cost for a *forward traversal* in a *path index*. Therefore, a *full index* has a higher average update cost than a *path index*. When n is small, since the cost for a *backward traversal* in a *nested index* is lower than the cost for the updated indices in a *full index*, a *nested index* has a lower average update cost than a *full index*. On the other hand, when $n \geq 5$, a *full index* has a lower average update cost than a *nested index*.

Fig. 10 shows the storage costs for a *full index*, a *multiindex*, a *nested index* and a *path index*, where $n = 3$, $m(1) = m(2) = m(3) = 1$, $K(2) = K(3) = 10$, and $K(1)$ varies from 1 to 50. Obviously, the *full index* can reduce the retrieval cost at the cost of increasing the storage cost; therefore, the *full index* has a higher storage cost than all the others. When $n = 3$, there are six indices in the *full index*, but there are three indices in the *multiindex* and there is only one index in the *nested index* and the *path index*. Since a *path index* records all the instantiations ending with the object at the end of the path and some instantiations may share some references, a *path index* has a certain degree of redundancy and has a higher storage cost than a *multiindex*. Moreover, when $K(i)$ is high enough, the *path index* may have a higher storage cost than the *full index*. Fig. 11 shows the storage costs for a *full index* and a *path index*, where $n = 3$, $m(1) = m(2) = m(3) = 1$, $K(2) = K(3) = 1$, and $K(1)$ varied from 50 to 80. From this figure, we can find that when $K(1) \geq 57$, the *path index* has a higher storage cost than the *full index*.

5. OPTIMAL INDEX CONFIGURATION

To reduce the high update cost for a long given path, we can split the path into several subpaths and allocate a separate index to each subpath [7,14]. Given a path, the number of subpaths and the index organization of each subpath define an index configuration. But the

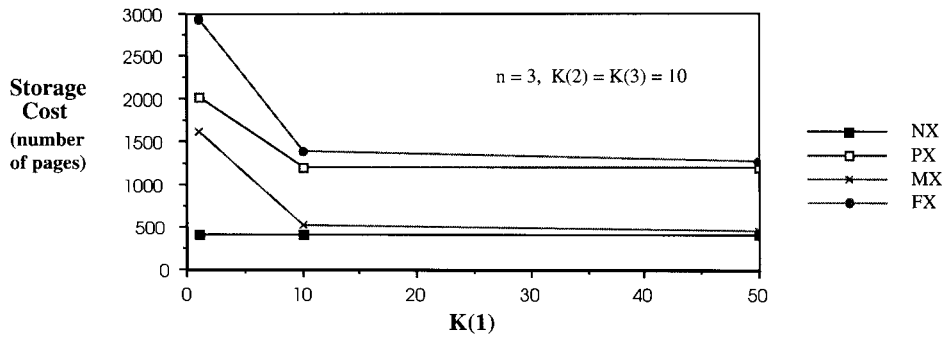

Fig. 10. The storage cost for $K(2) = K(3)=10$ under different values of $K(1)$.

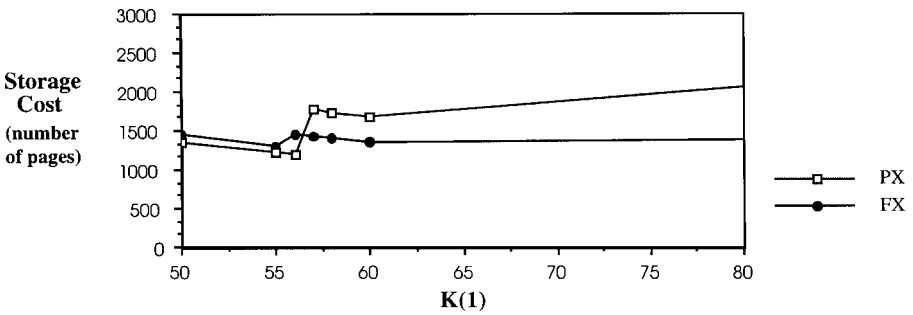

Fig. 11. The storage cost for $K(2) = K(3) = 1$ under different value of $K(1)$.

increase of the number of indices for subpaths will also increase the retrieval cost of scanning a number of indices, which results in a high retrieval cost. Since a low retrieval cost and a low update cost are always a trade-off in index organizations, we can establish a cost formula to look for a compromise between these two requirements. In [7], the authors proposed cost formulas to evaluate the costs of various index configurations. However, they do not support *partial instantiations* and do not consider more than one nested predicate along the path. In this section, we will propose cost formulas to determine the index configuration which can provide the best performance for various applications by taking into account various types of queries along a given path and a set of queries with more than one nested predicate along a given path.

For example, suppose that we have a path = $C(1).A(1)....A(3)$, where path-length = 3. This path can be split in several different ways. All the possible way to split this path s_i ($1 \le i \le 4$) are grouped as follows, where P_i^j denotes the jth subpath in the ith split way, and on the subpath P_i^j, $P_i^j\_c$ and $P_i^j\_a$ denote the starting class and the ending attribute, respectively:

Let us consider the split way s_3; we can allocate an index organiza10tion such as a *multiindex*(*MX*), a *nested index*(*NX*), a *path index*(*PX*) or a *full index*(*FX*), on each subpath. For example, $\{FX->P_3^1, NX->P_3^2\}$ denotes that a *full index* is allocated on subpath P_3^1, and that a *nested index* is allocated on subpath P_3^2, respectively.

Split Way	Subpaths	Boundary
s_1 :	$P_1^1 = C(1).A(1)$	$P_1^1\_c = 1 : P_1^1\_a = 1$
	$P_1^2 = C(2).A(2)$	$P_1^2\_c = 2 : P_1^2\_a = 2$
	$P_1^3 = C(3).A(3)$	$P_1^3\_c = 3 : P_1^3\_a = 3$
s_2 :	$P_2^1 = C(1).A(1).A(2)$	$P_2^1\_c = 1 : P_2^1\_a = 2$
	$P_2^2 = C(3).A(3)$	$P_2^2\_c = 3 : P_2^2\_a = 3$
s_3 :	$P_3^1 = C(1).A(1)$	$P_3^1\_c = 1 : P_3^1\_a = 1$
	$P_3^2 = C(2).A(2).A(3)$	$P_3^2\_c = 2 : P_3^1\_a = 3$
s_4 :	$P_4^1 = C(1).A(1).A(2).A(3)$	$P_4^1\_c = 1 : P_4^1\_a = 3$

In general, given a path $C(1).A(1)....A(n)$, the set of ways of splitting S is $\{s_1, s_2, ...s_r\}$, where r denotes the number of ways of splitting the path, and each way of splitting s_i ($1 \le i \le r$) is $\{P_i^1, ..., P_i^g\}$, where g denotes the number of subpaths of s_i. The organization sets for s_i are $ITS_i = \{IT_i^1 -> P_i^1, ..., IT_i^g -> P_i^g\}$ where $IT_i^k \in \{MX, NX, PX, FX\}$, $1 \le i \le r$ and $1 \le k \le g$.

5.1 Access Cost

Given a path $C(1).A(1)....A(n)$, a wayof splitting s_t and its index organization set ITS_t ($1 \le t \le r$), the retrieval cost for a query with a path-expression $C(i).A(i)....A(j)$ = '$O(j+1)$' ($1 \le i \le j \le n$) is denoted as $Cost\_A_t(i,j)$ and is obtained as follows:

$$\begin{aligned} Cost\_A_t(i,j) = {} & RC_t(i, P_t^l\_c) + \\ & + RC_t(P_t^l\_c, P_t^l\_a) \\ & + RC_t(P_t^{l+1}\_c, P_t^{l+1}\_a) + ... \\ & + RC_t(P_t^m\_c, P_t^m\_a) \\ & + RC_t(P_t^m\_a, j), \end{aligned}$$

where $P_t^{l-1}\_c < i \le P_t^l\_c, P_t^m\_a \le j < P_t^{m+1}\_a, 1 \le l \le m \le g$ and g is the number of subpaths of s_t. Moreover, $RC_t(a,b) = RC_a^b$, when $IT_t^y =$ 'FX', $P_t^y\_c \le a \le b \le P_t^y\_a$ and $1 \le y \le g$; that is, a *full index* is allocated on subpath P_t^y. If IT_t^y is 'MX', 'NX' or 'PX', $RC_t(a,b)$can be obtained as described in [3].

5.2 Maintenance Cost

Given a path $C(1).A(1)....A(n)$, a way of splitting s_t and its index organization set ITS_t ($1 \le t \le r$), the update cost for $O(i+1)$ of $O(i)$ ($1 \le i \le n$) is denoted as $Cost\_U_t(i)$ and is obtained as follows:

$$Cost\_U_t(i) = U_t^l(i), \qquad P_t^l\_c \le i \le P_t^l\_a.$$

Moreover, $U_t^l(i) = U_i$, when IT_t^l = '*FX*' and $1 \le l \le g$; that is, a *full index* is allocated on subpath P_t^l. If IT_t^l is '*MX*', '*NX*' or '*PX*', $U_t^l(i)$ can be obtained as described in [3]. Similar to the update cost, the deletion cost $Cost\_D_t(i)$ and the insertion cost $Cost\_I_t(i)$ can be easily obtained.

5.3 Cost Formulas

The total cost of an index configuration with a way of splitting s_t and its a corresponding organization set IT_t for a query with a path-expression $C(i).A(i)....A(j)$ = '$O(j + 1)$' ($1 \le i \le j \le n$) is denoted as

$$\alpha Cost\_A_t(i,j) + \sum_{m=1}^{n} \beta_m Cost\_U_t(m) + \sum_{m=1}^{n} \gamma_m Cost\_D_t(m) + \sum_{m=1}^{n} \delta_m Cost\_I_t(m),$$

where $\alpha + \sum_{m=1}^{n} \beta_m + \sum_{m=1}^{n} \gamma_m + \sum_{m=1}^{n} \delta_m = 1.$

Moreover, the total cost of an index configuration with a way of splitting s_t and its a corresponding organization set IT_t for a set of queries with path-expressions $C(i_1).A(i_1)....A(j_1)$='$O(j_1+1)$',$C(i_2).A(i_2)....A(j_2)$ ='$O(j_2+1)$',...., $C(i_k).\ A(i_k)$, ... $A(j_k)$ ='$O(j_k+1)$'($k \ge 1$, $1 \le q \le k$ *and* $1 \le i_q \le j_q \le n$) is denoted as

$$\sum_{q=1}^{k} \alpha_q Cost\_A_t(i_q, j_q) + \sum_{m=1}^{n} \beta_m Cost\_U_t(m) + \sum_{m=1}^{n} \gamma_m Cost\_D_t(m) + \sum_{m=1}^{n} \delta_m Cost\_I_t(m),$$

where $\sum_{q=1}^{k} \alpha_q + \sum_{m=1}^{n} \beta_m + \sum_{m=1}^{n} \gamma_m + \sum_{m=1}^{n} \delta_m = 1.$

Therefore, given a path $C(1).A(1)....A(n)$ and a set of queries with path-expressions $C(i_1).A(i_1)....A(j_1)$='$O(j_1+1)$',$C(i_2).A(i_2)....A(j_2)$='$O(j_2+1)$',....,$C(i_k).\ A(i_k)....A(j_k)$ ='$O(j_k+1)$'($k \ge 1$, $1 \le q \le k$ *and* $1 \le i_q \le j_q \le n$) the optimal index configuration can be obtained by trying all possible split types combined with all possible index organization sets and then finding one which can provide the minimum cost.

5.4 Simulation Results

In this subsection, we will discuss several simulations to find an optimal index configuration for queries along a given path against a class-aggregation hierarchy. In all these simulations, we assumed along that n = 8, $N(1)$ = 200,000, $m(y)$ = 1, $K(y)$ = 2, ($1 \le y \le n$), $UIDL$ = 8, kl = 2, ol = 6, pp = 4, f = 218, d = 146 and PS = 4096. Since the cost for an insertion(or a deletion) is in proportion to the cost for an update, we only consider the retrieval and update costs by letting the probabilities of insertion and deletion be 0 in the following simulations.he retrieval probability (denoted as R) varies from 1 to 0; at the same time, the update probability (denoted as U) varies from 0 to 1.

Fig. 12 shows the optimal index configurations for a simple query with a path-expression $C(1).A(1).....A(8)$ = '$O(8)$'. Fig. 13 shows the optimal index configurations for 2-degree complex type queries with the same probability of being performed, the path-expressions of which are $C(1).A(1).....A(8)$ = '$O(8)$' and $C(1).A(1).....A(4)$ = '$O(4)$', respectively. Fig. 14 shows the optimal index configurations for 8-degree complex type queries with the same probability of being performed, the path-expressions of which are $C(1).A(1).....A(j)$ = '$O(j)$', $1 \le j \le 8$. Fig. 15 shows the optimal index configurations for four queries in a general set with the same probability of being performed, the path-expressions of which are

$C(1).A(1).....A(8)$ = '$O(8)$', $C(2).A(2).....A(5)$ = '$O(5)$', $C(3).A(3).....A(7)$ = '$O(7)$' and $C(4).A(4).....A(8)$ = '$O(8)$', respectively. Fig. 16 shows the optimal index configurations for queries in a general set with the same probability of being performed, the path-expressions of which $C(i).A(i).....A(j)$ = '$O(j)$', $1 \le i \le j \le 8$.

From these figures, in general, we can find that a *full index* on the path $C(1).A(1).....A(8)$ is an optimal index configuration when the retrieval probability is high. The reason is that a *full index* can support any type of query along a given path with a lower retrieval cost than all the other index organizations, and the update probability is so small that the update cost is negligible. Therefore, the path does not have to be split into several subpaths to reduce the update cost. As the retrieval probability decreases (that is, the update probability increases), the update cost is not negligable, which results in a need for the path to be split to reduce the update cost. And for each split subpath, a *full index*, a *nested index* or a *path index* is allocated according to the query status on this subpath. Since a *nested index* and a *path index* cannot support *partial instantiations*, a *full index* is always a good choice to for allocation on each subpath as shown on Figs. 14 and 16. When the update probability is high, a *multiindex* on the path $C(1).A(1).....A(8)$ is a good choice because of the low required update cost and the negligable retrieval cost.

6. CONCLUSIONS

In this paper, we have proposed a new index organization for evaluating queries, called a *full index*, where an index is allocated for each class and its attribute (or nested attribute) along the path. From the analysis results, we have found that a *full index* can

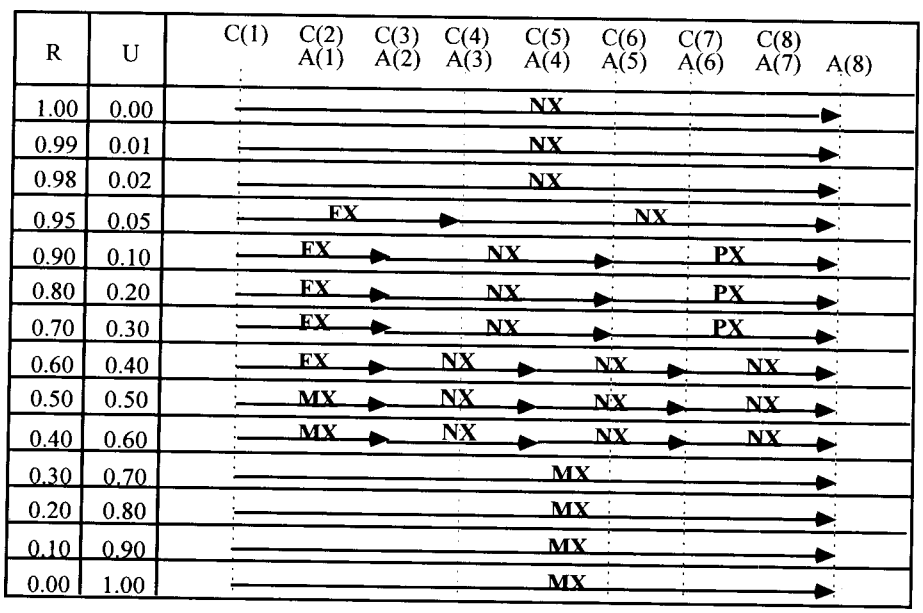

Fig. 12. The optimal index configurations for a query of a simple type.

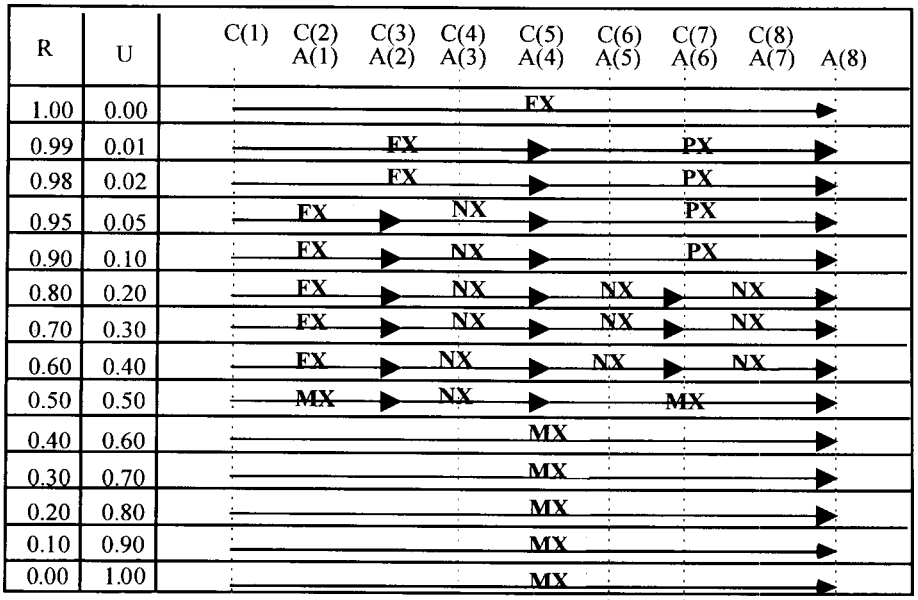

Fig. 13. The optimal index configurations for 2-degree complex type queries.

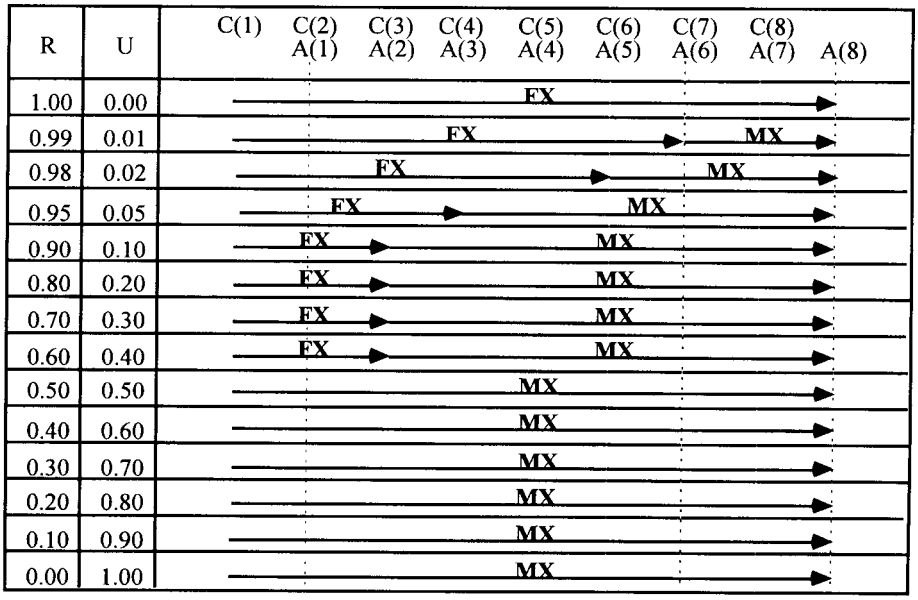

Fig. 14. The optimal index configurations for 8-degree complex type queries.

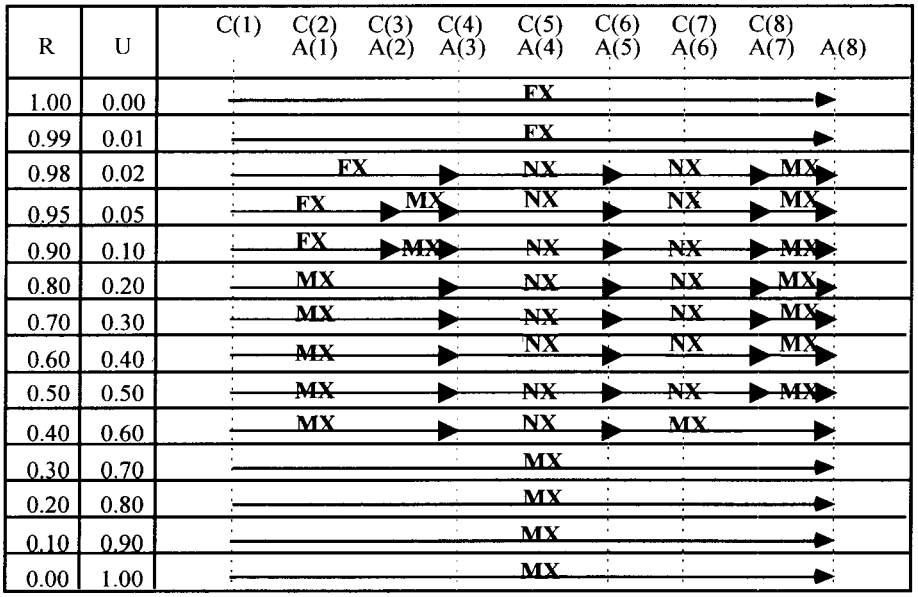

Fig. 15. The optimal index configurations for four queries in a general set.

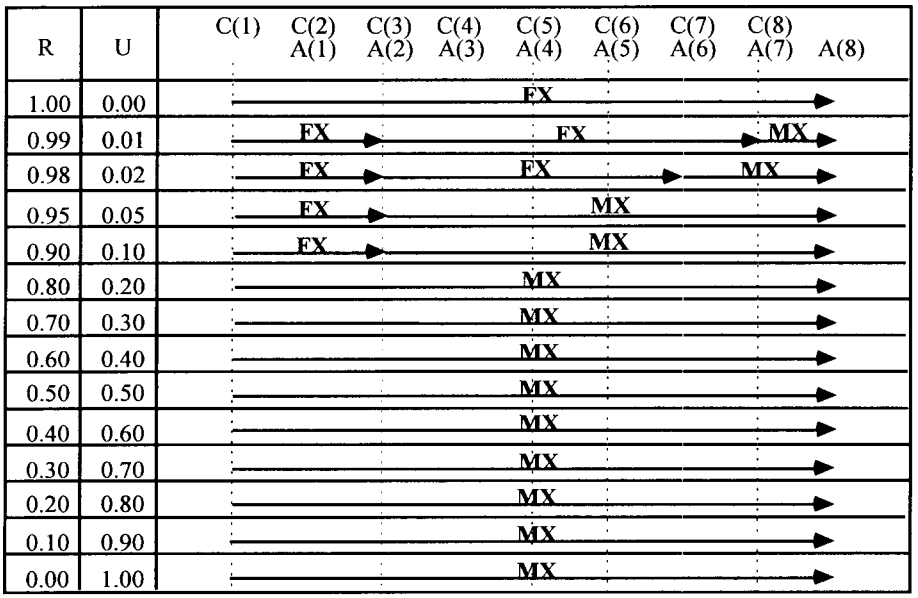

Fig. 16. The optimal index configuration for queries in a general set.

support any type of query along a path against a class-aggregation hierarchy with a lower retrieval cost than all the other index organizations (a *multiindex*, a *nested index* and a *path index*). Moreover, to reduce the high update cost for a long given path, we split the path into several subpaths and allocate a separate index on each subpath. Since a low retrieval cost and a low update cost are always a trade-off in index organizations, we have presented a cost formula which can be used to look for a compromise between these two requirements. In [7], the authors have proposed cost formulas which can be used to evaluate the costs of various index configurations. However, they do not support *partial instantiations* and do not consider more than one nested predicate along the path. In this paper, we have proposed cost formulas which can be used to determine the best index configuration to provide the best performance for various applications by taking into account various types of queries along a given path and a set of queries with more than one nested predicates along a given path. From the simulation results, in general, a *full index* is an optimal index configuration against a path when the retrieval probability is high. How to provide an efficient index organization for queries along more than one path against a class-aggregation hierarchy is a direction for future research.

REFERENCES

1. S. Abiteboul and R. Hull, "IFO: a formal semantic database model," *ACM Transactions on Database Systems*, Vol. 12, No. 4, 1987, pp. 525-565.
2. R. Bayer and E. McCreight, "Organization and maintenance of large ordered indexes, " *Acta Information*, Vol. 1, No. 3, 1972, pp. 173-189.
3. E. Bertino and W. Kim, "Indexing techniques for queries on nested objects," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 1, No. 2, 1989, pp. 196-214.
4. E. Bertino and L. Martino, "Object-oriented database management systems: concepts and issues," *Computer*, Vol. 24, No. 4, 1991, pp. 33-47.
5. E. Bertino, "An indexing technique for object-oriented databases," in *Proceedings of IEEE International Conference on Data Engineering*, 1991, pp. 160-170.
6. E. Bertino and L. Martino, *Object-Oriented Database Systems: Concepts and Architectures*, Addison-Wesley Publishing Inc., New York, 1993.
7. E. Bertino, "Index configuration in object-oriented databases," *Very Large Data Bases Journal*, Vol. 3, No. 3, 1994, pp. 355-399.
8. E. Bertino and P. Foscoli, "Index organizations for object-oriented database systems," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, No. 2, 1995, pp. 193-209.
9. R.G.G. Cattell, *Object Data Management: Object-Oriented and Extended Relational Database Systems*, Addison-Wesley Publishing Inc., New York, 1994.
10. S. Christodoulakis, J. Vanderbroek, J. Li, T. Li, S. Wan, Y. Wang, M. Papa and E. Bertino, "Development of a multimedia information system for an office environment, " in *Proceedings of the Tenth International Conference on Very Large Data Bases*, 1984, pp. 261-271.
11. D. Comer, "The ubiquitous B-tree," *ACM Computing Surveys*, Vol. 11, No. 2, 1979, pp. 121-137.
12. H. Ishikawa, F. Suzuki, F. Kozakura, A. Makinouchi, M. Miyagishima, Y. Izumida, M. Aoshima and Y. Yamane, "The model, language, and implementation of an object-oriented multimedia knowledge base management system," *ACM Transactions on Database System*, Vol. 18, No. 1, 1993, pp. 1-50.

13. A. Karmouch, L. Orozco-Barbosa, N. D. Georganas and M. Goldberg, "A multimedia medical communications system," *IEEE Journal on Selected Areas in Communications*, Vol. 8, No. 3, 1990, pp.325-339.
14. A. Kemper and G. Moerkotte, "Access support in object bases," in *Proceedings of the ACM International Conference Management of Data*, 1990, pp. 364-374.
15. A. Kemper and G. Moerkotte, "Access support relations: an indexing method for object bases," *Information Systems*, Vol. 17, No. 2, 1992, pp.117-145.
16. W. Kim and F. Lochovsky, "Indexing techniques for object-oriented databases," in W. Kim and F. Lochovsky (ed.), *Object-oriented Concepts, Databases, and Applications*, Addison-Wesley Publishing Inc., New York, 1989.
17. C. C. Low, B. C. Ooi, and H. Lu, "H-Trees: a dynamic associative search index for OODB," *ACM International Conference Management of Data*, 1992, pp. 134-143.
18. D. Maier and J. Stein, "Indexing in an object-oriented database," in *Proceedings of IEEE Workshop on Object-Oriented Data Base Management Systems*, 1986, pp. 171-182.
19. C. Meghini, F. Rabitti and C. Thanos, "Conceptual modeling of multimedia documents, " *IEEE Computer*, Vol. 24, No. 10, 1991, pp. 23-30.
20. E. Oomoto and K. Tanaka, "OVID: Design and implementation of a video-object database system," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, No. 4, 1993, pp. 629-643.
21. P. Valduriez, "Join indices," *ACM Transactions on Database Systems*, Vol. 12, No. 2, 1987, pp. 218-246.
22. D. Woelk, W. Kim and W. Luther. "An object-oriented approach to multimedia databases, " in *Proceedings of ACM International Conference Management of Data*, 1986, pp. 311-325.
23. D. Woelk, "Multimedia information management in an object-oriented database system, " in *Proceedings of the 13th Very Large Data Bases Conference*, 1987, pp. 319-329.
24. A. Yoshitaka, S. Kishida, M. Hirakawa and T. Ichikawa, "Knowledge-assisted content-based retrieval for multimedia databases," *IEEE Multimedia*, 1994, Vol. 1, No. 4, pp. 12-21.
25. Z. Xie and J. Han, "Join index hierarchies for supporting efficient navigations in object-oriented databases," in *Proceedings of the 20th Very Large Data Bases Conference*, 1994, pp. 522-533.

Chien-I Lee (李建億) was born in Taipei, Taiwan, R.O. C., 1965. He received the B.S. degree in computer science from Feng Chia University in 1987 and the M.S. degree in applied mathematics from National Sun Yat-Sen University in 1993. He received the Ph.D. degree in computer science from National Chiao Tung University in June 1997 and then joined the Institute of Information Education, National Tainan Teacher College, Tainan, Taiwan. He is currently an assistant professor, and his research interests include object-oriented databases, access methods, multimedia storage servers, video-on-demand, information retrieval and web databases.

Ye-In Chang (張玉盈) was born in Taipei, Taiwan, in 1964. She received the B.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1986, and the M.S. and Ph.D. degrees in computer and information science from Ohio State University, Columbus, Ohio, in 1987 and 1991, respectively.

Since 1991, she has been on the faculty of the Department of Applied Mathematics at National Sun Yat-Sen University, Kaohsiung, Taiwan, where she is currently a Professor. Her research interests include database systems, distributed systems, multimedia information systems and computer networks.

Wei-Pang Yang(楊維邦)was born on May 17, 1950 in Hualien, Taiwan, Republic of China. He received the B.S. degree in mathematics from National Taiwan Normal University in 1974, and the M.S. and Ph.D. degrees from National Chiao Tung University in 1979 and 1984, respectively, both in computer engineering.

Since August 1979, he has been on the faculty of the Department of Computer Science and Information Engineering at National Chiao Tung University, Hsinchu, Taiwan. In the academic year 1985-1986, he was awarded the National Postdoctoral Research Fellowship and was a visiting scholar at Harvard University. From 1986 to 1987, he was the Director of the Computer Center of National Chiao Tung University. In August 1988, he joined the Department of Computer and Information Science at National Chiao Tung University, and acted as the Head of the Department for one year. Then he went to IBM Almaden Research Center in San Jose, California for another one year as visiting scientist. From 1990 to 1992, he was the Head of the Department of Computer and Information Science again. His research interests include database theory, database security, object-oriented databases, image databases, and Chinese database retrieval systems.

Dr. Yang is a senior member of IEEE, and a member of ACM. He was the winner of the 1988 and 1992 AceR Long Term Award for Outstanding M.S. Thesis Supervision, the 1993 AceR Long Term Award for Outstanding Ph.D. Dissertation Supervision, and the winner of the 1990 Outstanding Paper Award of the Computer Society of the Republic of China. He also received the Outstanding Research Award of the National Science Council of the Republic of China.