UB-Tree Indexing for Semantic Query Optimization of Range Queries

S. Housseno, A. Simonet and M. Simonet

Abstract—Semantic query optimization consists in restricting the search space in order to reduce the set of objects of interest for a query. This paper presents an indexing method based on UB-trees and a static analysis of the constraints associated to the views of the database and to any constraint expressed on attributes. The result of the static analysis is a partitioning of the object space into disjoint blocks. Through Space Filling Curve (SFC) techniques, each fragment (block) of the partition is assigned a unique identifier, enabling the efficient indexing of fragments by UB-trees. The search space corresponding to a range query is restricted to a subset of the blocks of the partition. This approach has been developed in the context of a KB-DBMS but it can be applied to any relational system.

Keywords—Index, Range query, UB-tree, Space Filling Curve, Query optimization, Views, Database, Integrity Constraint, Classification.

I. INTRODUCTION

Performance enhancement is an important research area in the database domain especially when the DBMS deals with huge volumes of data. This problem has become crucial with the advent of applications/systems like Data Warehouses, Geographical Information Systems (GIS), spatial databases, multimedia databases, etc. Several techniques have been proposed and used to improve the performance of DBMS at the software level. Among these methods, there are data clustering, indexing data structures, query optimization, buffering, etc. Since a physical organization of data based on efficient indexing data structures with adapted query processing is one of the keys to efficient data retrieval, an important number of works in this domain have been proposed and implemented. There are two approaches to organize physically data on a secondary storage:

1) indexing based on a single attribute, e.g., hashing techniques, binary-tree, B-tree family [3], and
2) indexing based on multiple attributes, known as multidimensional indexing, e.g., multi-dimensional extensible hashing [7] R-trees [21], X-tree [8], Grid files [19], EXCELL [18].

This paper deals with the second approach. The indexing method presented in this paper was designed for an object data model that aims at unifying databases and knowledge bases [16]. This model has been implemented in the KB-DBMS prototype Osiris. As a DBMS, it is based on views defined by logical constraints on attributes; as a KBMS it performs instance classification on every object in the database.

However, the indexing method presented in this paper can be applied to any relational system provided it is possible to build a partitioning of the data (object) space into disjoint clusters.

In the Osiris KB-DBMS, a static analysis of the object data model enables the system to build a partitioning of the object space into disjoint blocks. Each block covers a portion of the object space. Instead of indexing directly the objects, the system indexes the blocks. For each query, the smallest set of indexing blocks that « contains » the query can be determined.

The problem addressed in this paper is that of representing the object space through the disjoint blocks, in order to support efficient access to the objects of a query. Blocks are by nature multi-dimensional. A block is a hyper-rectangle in a N-dimensional space, where N is the number of attributes. Each side of this hyper-rectangle represents an interval of the domain of its attribute. An Active block contains at least one object in the actual database. The set of Active blocks is indexed using a UB-tree [1], which is a multi-dimensional generalization of B-tree [3] based on Z-order curve [6].

The paper is organized as follows. A short survey about the multi-dimensional indexes and the type of supported queries is presented, then the Osiris system and its main concepts that are necessary to understand the partitioning approach of the object space are explained. The UB-tree indexing and how the object space (called Classification Space) is indexed using UB-trees organization is then presented. Finally, the processing of range queries in our approach is explained.

II. MULTI-DIMENSIONAL INDEXING

The indexing data structures which index data based on a single dimensional key like binary-tree, B-tree, etc. are efficient in database systems to support operations on data like retrieval, deletion, etc. However, these indexes are not suitable to situations where queries have multiple search keys [15], such as range queries and similarity queries, which play an
important role in many current situations such as Data Warehouses, spatial databases, multimedia databases, computer graphics, Geographical Information Systems (GIS), etc. To deal with these new database systems and applications; the representation of multi-dimensional data is an important issue.

Multi-dimensional data is to be seen as a collection of points (objects) in a higher dimensional space, i.e., whose dimension is greater than 1 [20]. For these object spaces, high-dimensional indexing methods have been considered as an important means to facilitate fast query processing. To support efficient retrieval in such high-dimensional databases, indexes are required to prune the search space.

Multi-dimensional indexes are required to support queries such as [15]:

1) Complete/Partial Range queries, and
2) Similarity queries:
   a) Similarity range queries: « find all objects in the database which are within a given distance from a given object », and
   b) K-nearest neighbor (KNN) queries: « find the K-most similar objects in the database with respect to a given object ».

This paper deals with complete range queries.

- Informally, a complete range query RQ is of the form « find all objects whose attribute values fall within a certain given range » [15]. For this type of query, a class C with n attributes can be considered as a n-dimensional space \(E_n\), defined as a Cartesian product of the domains \(D_1 \times D_2 \times \ldots \times D_n\), where the dimension \(D_i\) represents the domain of an attribute \(Attr_i\). In this space, an object \(o_j\) represented by the n-tuple \(\langle v_j1, v_j2, \ldots, v_jn \rangle\), represents a point in the \(E_C\) space and \(v_ji\) represents its coordinate in the dimension \(D_i\). In this space, a query is defined as: \(\{o \in E_C | o \in RQ\}\).

- Formally [15], if \(\delta_i\) is the range of a query along the dimension \(D_i\). The result of the query: \(Q=\{\delta_1, \delta_2, ..., \delta_m\}\); is the collection of \(\{o_i \in E_C\}\) that satisfy the condition \(v_j1 \in \delta_1, v_j2 \in \delta_2, ..., v_jn \in \delta_n\).

Multi-dimensional indexes such as R-trees [21] are not scalable in terms of the number of dimensions. When the dimensionality of data is high, the performances of R-tree-based index structures deteriorate rapidly [17]. Another type of indexing structures such as Grid files [19] and EXCELL [18] have been proposed. In this type of structure, data partitioning is dynamic, i.e., for each attribute, it is based on the distribution of the attribute values on its domain at a given moment.

To resolve the ‘dimensionality curse’ [15], [9] in these methods, some authors [5] have proposed to reduce the dimensionality of data by transforming data objects from a multi-dimensional space into one-dimensional space. Space Filling Curve is a way of mapping the multi-dimensional space into one-dimensional space [6]. A space filling curve imposes a linear order on the points by assigning an identifier to each one. A one-dimension index may then be used to index points (objects) by their identifier. As a result, the size of the indexed data is reduced, resulting in a smaller index size and a faster algorithm for search processing.

A space filling curve technique and one-dimension index are not used to index directly the objects. The indexing method presented in this paper uses them to index the disjoint blocks of the partition of the object space. In this approach, the partition of the object space is a semantic partition because it is based on the static analysis of object data model. This semantic partition provides an efficient query optimization because the query handles sets of objects instead of individual objects. In this type of approach, the phenomenon of partitions overlapping which happens with an index like R-tree is avoided. To explain how disjoint blocks are obtained from the object data model, a short presentation of main notions of Osiris is needed.

III. OSIRIS BASIC CONCEPTS

A full presentation of the Osiris system is not necessary to understand the semantic partition of the object space. The main notions that are useful for this purpose: P-Types, views, attributes and constraints are presented below.

Definitions

**P-Types.** The global object space is divided into disjoint sub-spaces where each sub-space, called a P-Type space, concerns the objects of a same family. « P » stands for the French « partagé » which means « shared ». As an example, the data model of the Information System of a car insurance company consists of three P-Types: the P-Type CAR, the P-Type CLIENT and the P-Type CONTRACT. In this data model, the object \(o_1\)\{name: Jack, age:40, sex: m, ClientId: 14524784AA, address:`01 Grande rue, Grenoble, FRANCE`,...\} belongs to the object sub-space of the P-Type CLIENT, the object \(o_2\}\{CarRegistration: 254 QDE 38, brand: Smart, doors number: 2, year: 2009,...\} belongs to the object sub-space of the P-Type CAR and the object \(o_3\}\{contracRef:14587515, InsuranceCoverages: Bodily Injury, BenefitID: 14524784AA\} belongs to the object sub-space of the P-Type CONTRACT. Fig. 1 shows a simple representation of this data model.

![P-Types Model](image)

Object sub-spaces

- **P-Types Model**
- **P-Type Client**
- **P-Type Contract**
- **P-Type Car**

Fig. 1 P-Types of the Car insurance company model
When designing an Information System in Osiris, choosing the P-Types of the application domain depends on the application needs. This is a designer’s decision.

P-Types are primitive concepts in the Description Logic paradigm, but not all primitive concepts are P-Types. The objects of a P-Type are meant to be shared by different categories of users, hence through different points of view, which are expressed by views in Osiris.

Views. A P-Type is organized as a hierarchy of views rooted in a minimal view that contains all the objects of the P-Type. A view is defined by the view(s) it specializes (except the minimal view, which is the root of the hierarchy), by its own attributes and by its own constraints defined on attributes, i.e., its own attributes and the inherited attributes from parent views.

Attributes. Attributes are defined within views. An attribute has a name and a type. The type of an attribute can be predefined (INTEGER, REAL, BOOLEAN, CHARACTER, STRING), a P-Type (i.e., a reference to a P-Type), and a collection (set, list) of a predefined type or a P-Type. Although attributes can be defined in any view (possibly in several views) of a P-Type, for the sake of simplicity we will consider in this paper that the attributes of a P-Type are defined in the minimal view and their domain is restricted by constraints in the views that constitute the P-Type.

Constraints. Constraints are Horn clauses whose literals are elementary Domain Predicates (in short DPs), i.e., predicates of the form Attr ∈ Domain, where Domain can be an interval (e.g., [10, 20]) or a set of enumerated values (e.g., {true, false}, {1, 3, 5, 7}, {blue, red, brown, yellow}).

Example. The P-Type PERSON is shown in Fig. 2 and Fig. 3 with very simple views.

```
view PERSON -- Minimal view of the P-Type PERSON
attr
name: STRING;
id: INT;
sex: CHAR in {m, f}; -- Domain constraint: Sex ∈ {m, f}
age: INT in [0..140]; -- Domain constraint: Age ∈ [0..140]
owns: set of CAR; -- The P-Type CAR is defined elsewhere
salary: INT ≥ 0; -- Domain constraint: salary ∈ [0..SUP]
…
age < 18 ⇒ salary < 1200,00
end PERSON;
```

view ADULT: PERSON -- Specializes the view PERSON
age ≥ 18 -- Domain constraint: age ∈ [18..140]
salary ≥ 600,00
end ADULT;
view SENIOR: ADULT
age ≥ 65
end SENIOR;
view CHILD: PERSON
age < 18 -- Domain constraint: age ∈ [0..18]
end CHILD;
view GIRL: CHILD
sex = f -- Domain constraint: sex ∈ {f}
end GIRL;
view BOY: CHILD
sex = m -- Domain constraint: sex ∈ {m}
end BOY;
view EMPLOYEE: ADULT
salary ≥ 1200,00
end EMPLOYEE;
view CEO: ADULT
salary ≥ 3000,00
end CEO;

Fig. 3 Description of the P-Type PERSON

The views and the P-Type defined above are very simple, in order to support the presentation of the Classification Space, which supports the indexing mechanism that is the basis of the semantic optimization mechanism.

Stable SubDomains. In a P-Type T, for each attribute Attr₁ let P(Attr₁) be the set of elementary predicates on Attr₁ that appear in all the assertions of all the views of T. Each elementary predicate has the form Attr₁ ∈ Dij where Dij is a subset of the domain of definition of Attr₁, i.e., Dij and j ∈ [1..NumSBD], where NumSBD is the number of the subset of the domain Dij.

An elementary predicate, i.e., Attr₁ ∈ Dij, determines a partition of Dj into two elements: d₁j and (Di - d₁j). The product of all the partitions defined by the predicates of P(Attr₁) constitutes a partition of Dj [14]. An element of this partition constitutes a block called Stable SubDomains (SSD), written d₁j. A subdomain is Stable because it verifies the stability property of an object with respect to the related attribute.

Definition 1: stability property of an attribute: When the value of an attribute Attr₁ of an object Oj varies within a SSD, e.g., dij ; j ∈ [1..NumSBD], the object Oj continues to satisfy exactly the same predicates of P(Attr₁).

A static analysis of the P-Type description allows determining the list of the SSDs of all the classifying attributes, see Fig. 4. Explaining the static analysis technique is outside the scope of this paper.

1 Classifying attributes are the attributes that take part in at least one domain constraint in a view of a data model.
P-Type description

Fig. 4 Static analysis of P-Type description and its output

Given the set of constraints defined in all the views of the P-Type PERSON, the products of the partitions for the attributes age, sex and salary lead to the following partitioning of their domain:

- **age**:
  - \( d_{11} = [0, 18] \)
  - \( d_{12} = [18, 65] \)
  - \( d_{13} = [65, 140] \)

- **sex**:
  - \( d_{21} = \{f\} \)
  - \( d_{22} = \{m\} \)

- **salary**:
  - \( d_{31} = [0, 600] \)
  - \( d_{32} = [600, 1200] \)
  - \( d_{33} = [1200, 3000] \)
  - \( d_{34} = [3000, \text{SUP}] \)

**SSD** is the set of all the stable subdomains of the attribute Attr. In the considered example, the SSDs of the attributes are:

- **SSD_age** = \( \{d_{11}, d_{12}, d_{13}\} \)
- **SSD_sex** = \( \{d_{21}, d_{22}\} \)
- **SSD_salary** = \( \{d_{31}, d_{32}, d_{33}, d_{34}\} \)

**Validity of SSD for a view**: A stable subdomain is valid for a view \( v \) if and only if:
1. It is valid for its parent views
2. It satisfies its constraints.

**Validity of a SSD for a P-Type**: A stable subdomain is valid for a P-Type \( T \) if it is valid for the minimal view of \( T \).

**Eq-class**: The Classification Space is a subset of the Cartesian product of SSDs of all the classifying attributes of the P-Type:

\[
\text{SSD}_1 \times \text{SSD}_2 \times \ldots \times \text{SSD}_j \times \ldots \times \text{SSD}_n = \{d_{1j}, d_{2j}, \ldots, d_{nj}\} \\
\text{where} \quad 1 \leq j \leq N, \text{where} \quad N \text{ is the number of classifying attributes.}
\]

The Classification Space is a N-dimensional space where each element, called Eq-class (for Equivalence Class) is a hyperrectangle represented by a N-uple of stable subdomains, i.e., \( <d_{11}, d_{2j}, \ldots, d_{nk}> \). See Fig. 5.

For the graphical representations, we limit ourselves to the 3D space. Thus, considering only the three attributes age, sex and salary, the Classification Space of the P-Type PERSON is represented in Fig. 6.

**Validity of an Eq-class for a view**: an Eq-class is valid for a view \( v \) if all the SSDs of its N-uple are valid for this view.

**Validity of an Eq-class for a P-Type**: an Eq-class is valid for a P-Type if all the SSDs of its N-uple are valid for the P-Type.

The valid Eq-classes of a P-Type PERSON are represented in bold on Fig. 6. For example, the Eq-class \( <d_{13}, d_{22}, d_{34}> \), that contains among others the object (age = 65, sex = m, salary = 4000) is valid because \( d_{13}, d_{22} \) and \( d_{34} \) are valid, whereas any object of the Eq-class \( <d_{11}, d_{22}, d_{33}> \) is invalid, because any person aged less than 18 (age \( \in d_{11} \)) can only satisfy \( d_{31} \) = [0..600] or \( d_{32} \) = [600..1200].
IV. INDEXING ACTIVE EQ-CLASSES IN OSIRIS

Before explaining the use of UB-trees in the indexing engine of the Osiris system, UB-trees [1] and the DRU algorithm [10] will be presented in this section.

Single-attribute indexing data structures are well tested and optimized. The need to index on many attributes and the emergence of multi-dimensional applications motivate the adaptation of single-attribute indexing data structures in these contexts. The transformation from multi-dimensional space to uni-dimensional space is an important and necessary step to use single-attribute indexing data structures. UB-tree indexing [1] is inspired by this approach.

UB-trees are a multi-dimensional generalization of B-trees [3] based on the Z-curve space filling curve [6].

A. Space-Filling Curve

The Space-Filling Curve is a method to map a multi-dimensional space into a one-dimensional space. In 1890, G. Peano was the first mathematician who constructed a curve that maps from the unit interval [0,1] to the unit square [0,1]^2 [11]. In 1891, Hilbert constructed a mapping of the whole space [12] and many curves have been proposed since [6].

Each curve has its own mapping function: Z order, Peano curve, Hilbert order, Gray order, U order, etc. Each curve visits the points of the multi-dimensional space one after another. The main difference between the curves is the choice of the next point to be visited. The multi-dimensional data universe is linearized to a one-dimensional space by representing a multi-dimensional point by its position on the curve. Consequently, the points are ordered, which permits to index them using a single-attribute indexing data structure, e.g., UB-trees. UB-trees are based on the Z-curve, which is presented in the next section. For other curves, see [6].

Space-Filling Z-curve

The mapping of a point from multi-dimensional space into one-dimensional space is done by calculating its position on the Z-curve, which is called its Z-value. Based on the binary representation, the Z-value is assembled by cyclically taking a bit from each coordinate of a point and appending it to those taken previously. Fig. 7

![Fig. 7 Bit-interleaving algorithm in N-dimensional space](image)

Fig. 7 shows how the Z-curve fills the two dimensional space.

The cost of Z-value construction is cheap and the work of [22] demonstrates that it has very good characteristics.

B. UB-tree

A UB-tree is a balanced multi-dimensional data structure based on the space filling Z-curve [6] and B-trees [3].

In a UB-tree, a Z-value, which is a position of a point on the Z-curve, is called a Z-address. Z-regions represent clusters of points in the indexed space. A Z-region is bounded by two Z-addresses which are the lower and the upper Z-addresses inside it. The Bounding UB-tree (BUB-tree) [4] does not index the Z-regions which do not contain objects (the dead space).

The UB-tree offers a hierarchical representation of space and also it partitions the whole space into a set of disjointive but consecutive Z-regions (Z-intervals). Each Z-region containing the indexed data is inserted into one leaf node in the UB-tree. On the other hand, the inner nodes contain super-Z-regions [10]. A super-Z-region bounds all the super-Z-regions in its subtrees.

The algorithms for insertion, deletion and point queries are similar to those implemented in B-trees except that the Z-address of the manipulated data must be computed before the execution of an algorithm. Due to the nature of range queries and the mapping into one-dimensional space, this query has its own algorithm in the UB-tree.

Range query processing

For range queries, the linear DRU algorithm proposed in [10] is used because its performance is better than the original linear algorithm proposed by Bayer-Markl [1], [2].

This algorithm is based on the intersection operation between the range query and the (super-)Z-regions mapped in the inner and leaf nodes. If a super-Z-region intersects the range query, so do its children. For more details, the reader is referred to [10].

The complexity of this algorithm is linear according to Z-address bit-length; i.e., O(n log(D)). A path stack is used to keep the current path being processed. The main steps of the DRU algorithm are:

1) Compute the Z-address of the query box lower and higher bounds, (Z_\text{lb}, Z_\text{hb}) respectively
2) Find the Z-region (leaf) which contains the Z_\text{lb}, set it as the current leaf and push its path onto the stack.
3) Search in the current leaf for data tuples that satisfy the range query.
4) If the lower bound of the right-neighbour-leaf Z-region is inside the range query, set it as the current leaf and goto 3.
5) The top of the stack is popped. It contains the parent node (node P) of the last treated leaf.
6) Peek node P to find an entry pointing to the next query-intersected node (node R). We have two cases:
   a. No such entry is found: remove node P from the stack and repeat step 6.
   b. One entry is found: retrieve the node R and push it onto the stack. If node R is a leaf, then goto step 3 otherwise repeat step 6.

In a two-dimensional space, Fig. 9 shows the super Z-regions and the Z-regions represented in the UB-tree. The tree itself was not represented for the sake of clarity. At each level, regions are larger but less numerous than those at the immediately lower level. The treatment of the RQ starts at the root level, where two super-Z-regions intersect RQ.

In the next level (level 1), three smaller super-Z-regions intersect the RQ. At level 2, four smaller super-Z-regions intersect the RQ. In the last level, which is the leaf level, four Z-regions intersect the RQ.

C. Using UB-trees in the Osiris indexing engine

In the indexing method, one UB-tree is used to index the active Eq-classes by their identifiers, instead of indexing directly the objects. This tree is called Active Eq-classes UB-tree (AEC UB-tree). The objects of each Eq-class are also indexed by another UB-tree, which is called Active Eq-class k Objects UB-tree (AECK-O UB-tree); k is the Eq-class identifier. A Z-region in an AECK-O UB-tree contains a set of Eq-class identifiers and pointers to the appropriate AECK-O UB-trees. A Z-region in an AECK-O UB-tree is a set of indexed objects (Fig. 10).

Recent systems have a large volume of RAM. Since the AECK-O UB-tree Z-regions contain a set of Eq-class identifiers and pointers, a whole AEC UB-tree and possibly the inner nodes of AECK-O UB-trees can be stored in the RAM. This is an efficient organization in the case of very large volumes of data.

V. QUERIES IN OSIRIS

For the insertion, deletion, and point query algorithms, firstly, the Z-address of the Eq-class of the object is calculated and then the UB-tree original algorithm is called. This paper deals complete range queries.

A. Range queries processing in Osiris

A query \( Q = \{\delta_1, \ldots, \delta_j, \ldots, \delta_n\} \) such that \( l_1 \leq \delta_1 \leq h_1, \ldots, l_j \leq \delta_j \leq h_j, \ldots, l_n \leq \delta_n \leq h_n \) can be seen as a hyper-rectangle in the N-dimensional space. This hyper-rectangle is bounded by a lower bound point \( P_l \) and an upper bound point \( P_u \) such that \( P_l = (l_1, l_2, \ldots, l_n) \) and \( P_u = (u_1, u_2, \ldots, u_n) \). In Osiris, these two points are transformed into the Z-addresses of Eq-classes.

To illustrate how UB-trees are used in Osiris, a two-dimensional space is used, with the dimensions age and salary of the P-Type PERSON (Fig. 11). Considering these two dimensions, the Eq-classes \( (d_{11}, d_{33}) \) and \( (d_{11}, d_{34}) \) are excluded from the P-Type, because of the constraint \( \text{age} < 18 \Rightarrow \text{salary} < 1200 \). The set of possibly valid Eq-classes is surrounded by bold lines.

In a N-dimensional space, an Eq-class is designated by a N-tuple of SSDs. In the example shown in Fig. 11, the Eq-classes are written \( (d_{11}, d_{31}), (d_{11}, d_{32}), \ldots, (d_{13}, d_{34}) \). This is a bi-dimensional representation. To obtain a one-dimensional representation, each SSD is assigned a binary code that is unique for each attribute. For example:

\[
\begin{align*}
\text{age} & \quad d_{11} = 00, d_{12} = 01, d_{13} = 10 \\
\text{salary} & \quad d_{31} = 00, d_{32} = 01, d_{33} = 10, d_{34} = 11
\end{align*}
\]

Applying the Z-order or (bit-interleaving), the unique
identifiers associated with the Eq-classes are computed. They are called Eq-class z-addresses:

\[(d_{11}, d_{31}) = 0000, (d_{11}, d_{32}) = 0001 \]
\[(d_{12}, d_{31}) = 0010, (d_{12}, d_{32}) = 0011, (d_{12}, d_{33}) = 0110, (d_{12}, d_{34}) = 0111 \]
\[(d_{13}, d_{31}) = 1000, (d_{13}, d_{32}) = 1001, (d_{13}, d_{33}) = 1100, (d_{13}, d_{34}) = 1101 \]

VI. CONCLUSION

The indexing method presented in this paper is designed for the P-Type object data model that aims at unifying databases and knowledge bases [13]. This model has been implemented in the KB-DBMS prototype Osiris. As a DBMS, it is based on views defined by the views they specialize, their own attributes and logical constraints on attributes. A static analysis of the object data model enables the system to partition the object space into a so-called Classification Space, whose elements are no longer individual objects but Equivalence Classes, named Eq-classes. The Classification space is used to optimize integrity checking, object classification, and for primary object indexing.

Moreover, in order to provide an efficient access to the objects referenced by an Eq-class, the UB-tree structure is used, which is a multi-dimensional generalization of B-trees based on the Z-curve space filling curve.

Our approach can be applied to any relational system by taking into account the integrity constraints of the database as a basis for the determination of SSDs and Eq-classes.

The indexing method presented in this paper is an efficient organization in the case of very huge volume of data. The processing of the DRU algorithm on the AEC UB-tree eliminates an important number of objects which are not inside the range query.

At present, the performance of the indexing method based on UB-trees is compared and analyzed with the performance of different indexing methods used in DBMSs.

VII. REFERENCES


