December 1993

Clustering Objects for OODBs in a Multiple Relationship Environment

Steven Tu  
Massachusetts Institute of Technology

Daniel Buehrer  
National Chung Cheng University

Follow this and additional works at: http://aisel.aisnet.org/pacis1993

Recommended Citation
http://aisel.aisnet.org/pacis1993/28

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 1993 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Clustering Objects For OODBs in a Multiple Relationship Environment

Steven Yi-Cheng Tu
Composite Information Systems Laboratory
E 53-315 Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139 U.S.A.

Daniel J. Buehrer
Institute of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, R.O.C.

Abstract

When data objects in databases are involved in more than one kind of relationships, we say that these objects are in a multiple relationship environment. Object-Oriented databases are an example of such a multiple relationship environment. That is, an object in OODBs may participate in n-to-1, n-to-n, 1-to-1, 1-to-n, or part-of and other relationships. The main goal of this paper is to take into account of all of these relationships when clustering objects in secondary storage. We develop a distance metric for objects to measure the degree of interrelationship, and apply a greedy algorithm to order the objects. After this sequence is constructed, we then can easily get this object sequence onto the secondary storage and achieve the purpose of clustering. We also discuss in this paper the features of our method that make it different from the previous proposed methods.

1. Introduction

1.1 Overview of Clustering in OODBs

The Object-Oriented (OO) paradigm has been recognized as a useful approach in knowledge representation, programming languages, and software engineering. Besides, the OO concept has also been applied to the database systems recently because of its extensibility and semantic richness. Object-Oriented databases (OODBs) offer the next-generation database applications: a good solution to the major shortcomings of traditional relational database systems. These are several characteristics of OODBs which we consider make it especially suitable for the newly-emerging applications, such as GIS, CASE, and CAD/CAM systems. These characteristics include: 1) the support of complex objects that enable users to define new complex data types instead of only system-defined flattened primitive data types; 2) the inclusion of abstract data types which make it easier to describe the dynamic behavior of the data, thus increasing the level of abstraction; 3) the use of inheritance that facilitates the database designers to reuse the attributes/code definitions; and 4) the support of object identity that makes it more natural to model the real-world entities.

Many OO languages such as Smalltalk and CLOS have been extended with the functions of persistence and shareability in order to make them become database programming languages. These OO languages have both the advantages of object-orientation and database systems. However, in order for these Object-Oriented database systems to be more practical and useful, some other database features must be added [9][23][24]. For example, a declarative query language [18], concurrency management [14], authorization [12], and storage control [13] must also be reconsidered in OODBs. Especially, the functionality of managing large amount of data safely and efficiently is crucial for OODBs design. In this paper, we concentrate on discussing and proposing an approach for clustering in OODBs.

Clustering is an important issue in the database design [13]. It is concerned with how to store the related data onto secondary storage as closely as possible in order to expedite the efficiency of data retrieval. Therefore, clustering can improve the performance of the database systems, which is the major problem of OODBs. In traditional relational database systems, each relation is independent. The relations are joined dynamically according to the later-relationship of relations when a user's query is processed. Consequently, it is sufficient to use either the n-ary storage model (NSM) or the decomposition storage model (DSM) [13][14] for the relational database clustering. But in OODBs, clustering is complicated by the multiple relationships among the objects. That is, an object can be both an instance of a class and a component of a complex object. Besides these aggregation and generalization relationships, other structural relationships such as versions and configurations [1] can also exist. In this situation, the objects are in a multiple relationship environment. This environment affects the clustering methods that can be used because we have multi-dimensional goals while storing the related objects onto the single-dimensional storage. In order to solve this problem, we propose a method in this paper for compromising on the possible relationships among objects and then ordering the objects into a clustering sequence. After the sequence is constructed, we can then easily map the sequence into the secondary storage and preserve the multiple relationships of the objects, thus achieving the purpose of clustering.

1.2 Previous Related Work

Most previous papers on OODb clustering have been focusing on complex objects (nested objects). In [2] Banerjee and Kim studied the possible universal methods of a DAG (Directed Acyclic Graph), which is in fact the complex object hierarchy. In [3] Davis discussed the representation of large complex objects and the algorithms to dynamically insert and delete objects in the large complex objects with an emphasis on optimizing performance. Another complex object storage model which is not based on the the object reference relationships but rather on the object identifiers was presented in [22]. But the above research only restricted their discussions to a single relationship between objects and did not cover the thorough semantics of the OO paradigm. [1] was the first paper to take into account the multiple relationships (Inheritance, Version, Configuration) in OODBs, and a smart clustering algorithm was also given that
could exploit these structural relationships for clustering objects. However, only one kind of relationship could be input into the smart algorithm instead of all possible multiple relationships. The actual pioneering paper dealing with the clustering techniques in a multiple-relationship environment is [5]. In [5], related objects can be connected into a multi-graph by assigning a weight according to each relationship between two objects. Then, a level clustering algorithm is used to construct a maximal spanning tree from this weighted multi-graph. The clustering problem and our approach in this paper must resemble [5]. However, our approach is different from [5] in that we actually develop a metric to measure the distance between objects, which is a concrete calculation of the weight wij in [5]. Besides, the distance metric in our approach also takes into account the cardinality (size) of each relationship set explained later, which was not considered in [5].

The remaining sections of this paper are organized as follows: In Section 2 we give an OODB example model in order to clarify our problem and also pave the way for future discussions. Section 3 presents our proposed approach, including the distance metric and sequencing algorithm. In Section 4, we discuss the features of our approach and also give some conclusions.

2. The Example Model and Clustering Options

In order to better explain the clustering options in a multiple relationship environment and use this example as an illustration when we introduce our approach, we borrow a sample model from [8]. We choose this example for three reasons. One reason is that most of the important concepts of OO are incorporated into ORION [19], which make it a perfect example to be discussed in this paper. Another reason is that most of the major issues of ORION, including the query model [18], schema evolution [19], and transaction control [14] are already discussed in the literature. The third reason is that ORION system is not only a proposed model but also implemented and operational [20]. The following graph is a schema example of a multi-table in an OODB. Each square box is a class with attributes for that class defined inside the box. The dashed line represents the inheritance (is a) relationship and the solid line represents the aggregation (part of) relationship. The part of relationship is the inter-object relationship of the above-mentioned complex object relationship, and the is a relationship says that each instance of a subclass is also an instance of its superclasses. Another relationship which is not explicit in this example is the instance of relationship. All of the instances belonging to the same class are considered to possess this relationship. In this paper, if only one kind of relationship was considered, we call that clustering in a single relationship environment, and if all relationships are considered, then clustering is in a multiple relationship environment.

Several clustering options for the above sample database may be adopted. Option 1 is to cluster all objects which belong to the same class. This involves storing objects according to the instance-of relationship. For example, all objects in Vehicle would be stored together. Option 2 is to cluster all objects belonging to the same class hierarchy. This involves storing objects according to the is-a relationship. For example, all objects in Vehicle, Automobile, and Track would be stored together. Option 3 is to cluster objects that other objects inter-reference. This involves clustering objects according to the the part_of relationship. For example, the objects in Vehicle, Company, and Employee would be stored together. It should be noted that the above clustering options are just some of the possible relationships in OODBs. Other more complicated relationships from application semantics might also exist. For example, Figure 2 represents the version relationship for the sample database.

![Figure 1: A sample database](image)

Figure 1: A sample database

![Figure 2: A version network](image)

Figure 2: A version network

Each clustering option has major effects on the retrieval performance for the application's access pattern. The access pattern is defined as the traversal sequence for the classes involved in the user's query. We list several possible user's queries to the sample database. A class variable associated with a star (*), like Vehicle, means to retrieve from this class and all of the subclasses rooted at that class.

Query 1: (Vehicle select * (V color = "blue"))

Query 2: (Vehicle select V : (V drivetrain transmission = "fuel" and V manufacturer name = "Ford"))

Query 3: (Vehicle select N : (V weight > 1000 and color = "red"))

Query 4: (Vehicle select N : (V color = "blue")) and (-V manufacturer name = "Ford")

For each query above, there is a corresponding clustering option that makes the retrieval performance of that query more efficient. For example, the clustering option 1 favors query 1. Similarly, clustering option 2 favors query 2 and clustering option 3 favors query 3. However, no clustering option that we mentioned above favors query 4 because it involves more kinds of relationships than other queries. In summary, each clustering option might be optimal for some kinds of access patterns (queries) but may sacrifice performance on other access patterns. Furthermore, dedicating the clustering to one kind of relationship will incur very poor access performance for other queries. For instance, clustering objects according to part of relationship in the same disk block will cause objects of any single class to be spread out over a larger number of pages than if the class is the only class to be stored in one block. However, the multiple relationship property is an inherent constraint for both Object-Oriented systems and next-generation applications. Therefore, some way to compromise all of these relationships while dealing with clustering must be developed in order to achieve a better average performance for users' ad hoc queries. This is the motivation for our method to solve this problem.

3. The Distance Metric and Clustering Algorithm
3.1 Outline of Our Approach

As mentioned before, we are going to use distance to measure the inter-relationship of objects. Each object, according to the kind of relationships in which it may participate, can be represented by an n-tuple of category 1, category 2, ... category n. If there are n possible relationships involved in the system. For example, an object of (C1, C0, V1) can be thought of as belonging to class C1, complex object C0, and version number V1. One problem is that most of the values of each category in the n-tuple are qualitative and discrete data. This means that to calculate the distance between two objects, some quantifying method must be used in order to convert the n-tuple into numeric data, from which a distance can be measured. There are two basic constraints which we consider must be satisfied if some quantifying method is to be used. The first constraint is that any quantifying configuration which faithfully represents the patterns existing among objects must guarantee that more related objects should be represented by points that are closer. The second constraint is that if two objects are non-comparable inter-related, then there should not be any false or greater distance comparison after the transformation. These two constraints can be represented as follows where $\Theta$ denotes the semantic relationships between objects and $\alpha$ denotes the closer than operator.

Constraint 1: $(\Theta_{01,02} \in \Theta_{01,03} \Rightarrow \text{Distance}(01,02) < \text{Distance}(01,03))$

Constraint 2: $\exists (\Theta_{01,02} \in \Theta_{01,03} \text{ or } (\Theta_{01,02} \notin \Theta_{01,03})) \Rightarrow \text{not } (\text{Distance}(01,02) < \text{Distance}(01,03) \text{ or Distance}(01,02) < \text{Distance}(01,03))$

Our original idea was to quantify each object from a qualitative n-tuple to a numeric n dimensional coordinate system and apply a multivariable analysis method such as the principle component analysis [7] to project these n dimensional points into the principle components, and then maximize the preservation of relationships from the original n dimension space. We did not adopt this approach because when we quantize each object into an n-dimensional numeric coordinate, we lose some semantics. This will not satisfy the second constraint. For example, if category 1 represents the class to which one object belong and we quantize class Vehicle as 1, Company as 2 and VehicleDriver as 3, then this makes Vehicle more related to VehicleDriver then to Company (1-3, 1-2, 2-3). However, this quantified transformation of the qualitative data is wrong. Therefore, we develop a metric to directly measure the distance between each two objects instead of transforming each object into an n-dimensional point. Our metric satisfies both the first and second constraints, thus does not lose any semantics of object inter-relationships.

3.2 Distance Metric

A relationship set is defined to be the set of objects that are constrained in the relationship and the size of the relationship set is defined to be the number of objects in this set. For example, one relationship set is the class Vehicle, and its size is the number of instances in Vehicle. Another relationship set is the complex object instance involved in Vehicle, Company, and Employee, and its size is the number of objects that comprise this complex object instance. In this paper, each relationship set will be denoted Rm and its size will be denoted |Rm|. We can now define the parameters and the distance metric, which is a measure of distance between each two objects, as follows.

$$\text{Distance}(\alpha, \beta) = \sum_{m=1}^{n} p_m \times (\bar{d}_{ij,Rm} \times (1/2) \times |Rm|) + \beta_{ij,Rm} \times (1/2) \times L \times s$$

$\alpha$: number of relationship sets
$\beta$: number of objects in the system
$s$: object size
$L$: total size of all objects (L = k \times s)
$b$: disk block size

The above formula is rather generalized since it encompasses every possible relationship in the system and is calculated by summing the effects of all relationships. In this formula, we assumed that the size for each object in the system is the same, and is equal to s. $p_m$ is the probability that the system is clustered by relationship Rm. This probability comes from the probabilistic distribution of application access patterns and can be specified by system designers. If some access pattern occurs more often or is more important, then its corresponding probability will also be larger. However, we assume that each probability $p_m$ is the same and is equal to 1/|m| in the discussion below. Now explain the meaning of this distance metric. It is the expected value of the object distance by considering all semantic relationships. This expected distance is formed by summing the product of the probability $p_m$ and the average distance (see Figure 3) of objects clustered by each possible relationship. The $Rm$ is the m-th relationship set or the property category m if objects are represented by an n-dimensional coordinate.

$\delta_{ij,Rm}$ and $\beta_{ij,Rm}$ are mutually-inverse characteristic functions that have values either 1 or 0. They can be stated as follows:

$$\delta_{ij,Rm} = 1 \text{ if object } i \text{ and object } j \text{ are in a relationship set } m$$
$$= 0 \text{ if object } i \text{ and object } j \text{ are not a relationship set } m$$

$$\beta_{ij,Rm} = 1 \text{ if } \delta_{ij,Rm}$$
$$= 0 \text{ if } \delta_{ij,Rm} = 0$$

Now assume the simple case that only instance-of and part-of relationships exist in the system. Suppose that the instance-of relationship is indexed as R1 and part-of relationship is indexed as R2. We show how the distance of each two objects of the example in Section 2 can be calculated. Because we assume that P1 and P2 are the same, they will be both equal to 1/2.

R1: 1/2 + 1/2 + 1/2 + 1/2 + 1/2 + 1/2
Case 1: 1/2 + 1/2 + 1/2 + 1/2 + 1/2 + 1/2
Case 2: 1/2 + 1/2 + 1/2 + 1/2 + 1/2 + 1/2
Case 3: 1/2 + 1/2 + 1/2 + 1/2 + 1/2 + 1/2
Case 4: 1/2 + 1/2 + 1/2 + 1/2 + 1/2 + 1/2

For example, the distance between two objects in class Vehicle will be 1/2(1/2 + |Vehicle| + 1/2 + 1/2) if they are not
cyclically referenced. However, if oi and oj are further recursively referenced [11], then the distance will be 1/2(1/2) in vehicle + 1/2 (COP) where COP is the complex instance that contains oi and oj. It can be shown that the distance for the above cases has the property that distance (case 1) < distance (case 3) < distance (case 2) < distance (case 4). This satisfies our constraint 1 because more related objects have less distance. However, the distance for case 3 and case 2 is non-comparable. It depends on the size of the relationship set ( 3 and 10). This also satisfies our constraint 2 because the closeness relationship is not transitive. If oi and oj are in the same class and oj and ol are in the same complex object, then an absolute semantic closeness can be stated for oi to oj and oj to ol. The two constraints will be satisfied because the expression |Rb| < 1 holds for every relationship set in the system. We now explain how the relationship size is included in the formula. This can be demonstrated in Figure 3. If two objects are in the same class and we also cluster by the instance of relationship, then the distance will be the average of the class size. Figure 3 explains how to calculate the distance between two objects. The distance is obtained for the case.

A similarity matrix [26] can then be constructed. This is a similarity matrix whose entry at each object is the distance between any two objects \( M(i,j) = \text{distance}(i, j) \). This matrix is symmetric with all diagonal elements equal to 0. From this matrix, we can then find the spanning path which includes all of these k objects by applying the clustering algorithm in the next section.

3.3 Sequencing Algorithm

Since we must eventually put all of the objects onto the linear secondary storage, a better way is to sort the objects of the system into a one-dimensional data structure. Then the mapping of the one-dimensional data structure to the secondary storage will be straightforward. The previous research in [5] constructed objects into a maximally weighted spanning tree. However, because a tree is a two-dimensional data structure, some traversal methods such as DFS or BFS must then be applied in order to linearize the tree. We consider this to be disadvantageous because when we linearize the spanning tree, some distances relationships, which are also semantic relationships of objects, may be compromised. Therefore, our linearizing algorithm optimizes the result as a spanning path, by which all of the multiple relationships can still be preserved to some degree. After the distance matrix is constructed by assigning the distance between each two objects, the following sequencing algorithm [11] can be used to order the objects into a path sequence. This input of this algorithm is the distance matrix which reports the weighted average of each relationship between each two objects. We also assume that there are totally k objects in the system. Each one of these k objects can be a member of more than one kind of relationship set. This algorithm is based on the greedy strategy because it works by selecting one object at a time into the current spanning path. This can be done by selecting an unprocessed object which adds the minimum distance to the current distance of the current path.

The problems of finding the optimal clustering scheme for generalised access patterns and the finding of shortest-spanning-path (STSP) are both NP-complete [15][17]. Our Totally Shorter Spanning Path (TSSP) algorithm was inspired by the Shorter-Spanning-Path (SSP) algorithm proposed in [11]. The main difference between TSSP and SSP is that in Step 3, TSSP randomly chooses an object from the remaining as the current spanning path at a particular position which adds the minimum distance but SSP iterates over all of the unprocessed objects in W to find the minimum added distance. The total distance of the resulting spanning path for TSSP is less than that of SSP because in TSSP the checking of all of the unprocessed objects in W also covers the checking of the randomly selected object from W in SSP. In order to clarify the difference between these three spanning path constructing algorithms we compare TSSP, SSP, and SSP with respect to time complexity and optimality in the following:

Optimality: TSSP > SSP > STSP
Execution time: STSP(O(k^3)) > TSSP(O(k^2)) > STSP(exponential)
Because the secondary storage access unit is a disk segment or a block, the spanning path must be cut into path segments according to the ratio of the object size to the block size (Ld). Notice that our clustering sequence is not interrupted by segmenting the spanning path. The change only occurs at the segmentation boundary. This can still be compensated for by placing the adjacent path segments into the same space and adjusting block sizes, which can save disk access time.

In the next section, we discuss the common features of clustering in database systems and compare our approach with the previous proposed methods.

4. Discussion of Our Approach and Conclusions

4.1 Discussion

In this section, we discuss several features of our approach, which are also the criteria for clustering. We only list those points that we consider different from previous proposed approaches.

1. Access performance: Clustering fastens the access performance of related data in the database. The related data are correlated by multiple kinds of relationships. Therefore, the access performance is the expected average access cost of all path segments, each of which retrieves data according to some relationship. Because there is a kind of relationships in the system, a kind of access patterns (queries) denoted Q1,Q2,..., and Qn may be issued to the databases. We also assume that each clustering schema based on relationship set in stores access pattern Qj,Rn. The retrieval cost of objects for Qj,Rn will be proportional to the distance of any two objects belonging to relationship set Rn. If all of the access patterns are considered, then the average retrieval cost for Q1,Q2,..., and Qn, which equals (P1\cdot Cost(Q1)) + P2\cdot Cost(Q2) + ... + Pn\cdot Cost(Qn), will be proportional to our distance metric, if the access probability of Qj,Rn is Pj. This means that our distance metric is also based on the expected average access cost for possible access patterns Q1,Q2,..., and Qn. Our approach gains in access performance in that we are constructing a totally shorter spanning path that sums up less total access cost. Therefore, our approach has a better chance to induce better average access performance. We want to emphasize that the optimal access performance is very hard to achieve because it is an intractable problem. The advantage of our approach comes from the merits of the shorter spanning path (TSP), which are shown in [11]. Furthermore, our sequencing algorithm results in a TSP which is better than TSP in [11].

2. Automatic and adaptive: Many DBMS adopt the scheme of receiving hints from users, which are frequent access patterns, and support a default strategy if hints are not specified by users. Two drawbacks might exist for this scheme. One is that users have to deal with the multiple relationships of objects, which are intrinsically multiply-dimensional and which could even conflict with each other. Users probably are able to specify one kind of relationship but are unable to combine all of the relationships and then give hints to the DBMS. The other problem is that sometimes users will need to specify a network of the related schema graph of a class for clustering [26]. For example, it may be useful for users to give hints to store physical components of a vehicle object in the same block, but not the Company objects for the manufacturer attribute of the vehicle objects. This problem comes from the fact that users have little knowledge of the extensional information, such as the class size, the actual object reference relationships and object sharing relationships, of the database. Therefore, a better approach is for the system to automatically cluster objects, because the system knows more information for clustering than the users. Our approach is automatic because the calculation of the distance metric and the sequencing process are all done by the system. However, our approach is also adaptive. This is from our sequencing algorithm Step 2 that when there are several chosen objects that all satisfy the condition (M(R1,S1) + M(R2,S2) + M(R3,S3) is minimized). Users can then be involved in which object should be chosen. This means our clustering method in adaptive when the system has alternatives to decide and the users help to make the decision.

3. Resulting: Clustering in databases is used to divide the database into homogenous groups and store each homogenous group in a multi-disk system or distributed environment in order to exploit the parallelism and to reduce access times. Our clustering algorithm outputs the result as a path sequence. This sequence can then be cut to form homogenous groups, which we call path segments above. Each path segment can be allocated to a single process or to each block in a multi-disk system and a parallel search techniques may be used to fetch the access speed for the users' queries [6].

4.2 Conclusions

A method of measuring distances between objects and a sequencing algorithm are proposed in this paper. Our distance metric covers the thorough semantic relationships of objects in the OODBs. This is important for Object-Oriented database clustering because the OODBs applications impose more complicated access patterns than traditional relational database applications. Any clustering schema for OODBs which is offered to versatile applications should not favor some specific access pattern but sacrifice others. A better average performance for each kind of access pattern might be achieved in order to generalize the use of the Object-Oriented database system. Our method is based on the simplicity of calculating distance but without losing flexibility to consider multiple kinds of object relationships. It is further enhanced by our sequencing algorithm, which finds a direct mapping between the object sequence and secondary storage. Because of the importance of improving the efficiency of OODBs for distributed systems, we feel that clustering in multiple relationship environments should remain a major research issue for some time to come.

References


