Performance Evaluation for Clustering Algorithms in Object-Oriented Database Systems

Jérôme Darmont
Ammar Attoui
Michel Gourgand

Université Blaise Pascal - Clermont-Ferrand II, Laboratoire d’Informatique,
Complexe scientifique des Cézeaux, 63177 Aubière Cedex, France

Abstract. It is widely acknowledged that good object clustering is critical to the performance of object-oriented databases. However, object clustering always involves some kind of overhead for the system. The aim of this paper is to propose a modelling methodology in order to evaluate the performances of different clustering policies. This methodology has been used to compare the performances of three clustering algorithms found in the literature (Cactis, CK and ORION) that we considered representative of the current research in the field of object clustering. The actual performance evaluation was performed using simulation. Simulation experiments we performed showed that the Cactis algorithm is better than the ORION algorithm and that the CK algorithm totally outperforms both other algorithms in terms of response time and clustering overhead.

Keywords: Clustering, Computer systems performance evaluation methodology, Object-oriented databases, Simulation.

1 Introduction

Clustering is a technique that is widely used to improve the performances of conventional Database Management Systems (DBMSs). Clustering means storing related objects close together on secondary storage so that when one object is accessed from disk, all its related objects are also brought into memory. Then access to these related objects is a main memory access that is much faster than a disk access. With the arrival of Object-Oriented Databases (OODBs), existing clustering algorithms (mainly in the field of relational databases) had to be adapted to the additional semantics (such as inheritance, etc.) introduced by the object-oriented data models. It appeared that a good object clustering was critical to the performance of OODBs [16]. The aim of this paper is to propose a methodology in order to compare the performance of the different clustering strategies that can be implemented in OODBs. Several methods can be used to evaluate the performances of a DBMS. Benchmarks generally propose a standard database and a series of operations that run on this database. Thus, performance measurement directly depends on the reactions of the tested DBMS. Several
benchmarks have been specifically designed for object-oriented databases, like the Synthetic Benchmark [15], the HyperModel Benchmark ([1], [4]), the O01 Benchmark [6] or the CluB-0 Benchmark [2]. However, some OODB designers or clustering algorithm authors prefer to use simulation ([8], [10], [13]), because simulation allows to specifically measure performance improvements due to one or another clustering policy. [17] proposes a dual performance evaluation method, performing simulations that use the database introduced by the CluB-0 Benchmark. One last way to determine the advantages of a given clustering method is mathematical analysis as it is performed by [7]. This approach is however limited because the obtained results are qualitative rather than quantitative and sharp performance criteria cannot be extracted. The aim of the approach we suggest to use is to propose a modelling methodology that allows to compare OODBs' performances, and especially clustering strategies' performances. Modelling may lead either to simulation or to the application of exact analytical methods whenever possible. We applied our methodology to three object-oriented clustering algorithms that are representative of the current research in the field of OODBs: Cactis [14], CK [9] and ORION ([3], [15]). The main advantage of our approach opposed to the use of benchmarks is that it allows, by providing a common environment, to specifically compare clustering algorithms, in a way that is totally independent of any environment associated with the DBMSs that implement the clustering algorithms we intend to compare. For instance, physical storage methods and buffering strategies also influence the DBMS' global performance. Furthermore, our approach also allows to a priori study the behaviour of algorithms (like CK) that are not implemented in any DBMS. Thus we can compare their performances to those of already implemented algorithms. This paper is organised as follows. We start by presenting the modelling methodology we used. Section 3 is dedicated to our study: we apply our modelling methodology to obtain a knowledge model and an action model. Then we present in Section 4 the three studied clustering algorithms. The simulation results are given in Section 5. We end this paper with a conclusion and a brief discussion about future research directions.

2 Modelling Methodology

OODBs are complex systems. Modelling their behaviour may as well be a complex task. This is the reason why we propose an approach dedicated to the study of such systems. This modelling approach carry out a model according to an iterative process [12]. This process is divided into four phases:

- **Phase 1**: analysis and formalizing of data, this system specification leads to the design of the knowledge model; it is a crucial step in the modelling process;
- **Phase 2**: translation of the knowledge model into an action model using a formalism allowing its exploitation to provide performance criteria;
- **Phase 3**: exploitation of the action model to provide performance criteria;