A Novel Query-Driven Clustering-Based Technique for Vertical Fragmentation and Allocation in Distributed Database Systems

Adel A. Sewisy, Faculty of Computers and Information, Assiut University, Assiut, Egypt
Ali A. Amer, Science College, Taiz University, Taiz, Yemen
Hassan I. Abdalla, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

ABSTRACT

In this paper, heuristic query-driven clustering-based vertical fragmentation technique is efficiently developed. The intrinsic idea is to meet the idealistic case of DDBS design which comes to happen as each query attune “closely match” its relevant fragment. The proposed technique is mainly sought to breed clusters of queries in the first place. Consequently, these clusters would be tacitly used to generate intended disjoint fragments. Moreover, the allocation process has been considered so that replicated and non-replicated scenarios of data are applied. This technique basically meant to be efficaciously applicable at the initial stage of DDBS design without the need for data statistics or empirical results, in either dynamic or static DDBS environment. Many existing design-related techniques are being incorporated to make a promising work, particularly as communication costs being the foremost design objective. Throughout this work, the experimental results and internal evaluation are extensively illustrated to demonstrate the effectiveness and validity of proposed technique.

KEYWORDS
Allocation, Clustering, DDBS, Model, Query Clustering, Replication, Vertical Fragmentation

1. INTRODUCTION

Throughout the past few decades, the topic of DDBS design has been thoroughly investigated. As a matter of fact, this comes down to the huge impact of proper design on DDBS rendering. For DDBS design, the most important challenges to be dealt with are fragmentation and allocation. Ostensibly, the fragmentation and allocation methods fundamentally seek to minimize the amount of irrelevant data transmission among different sites as distributed query processed.

As a result, this work aims at finding an accurate fragmentation and allocation technique to promote the system throughput through significantly dwindling communication costs. Meanwhile, it is worth referring that this work reconciles with saying of (Hammer and Niamir, 1979) and (Navathe, Ceri, Wiederhold and Dou, 1984) that the general vertical fragmentation dilemma is “heuristic in nature,” and that optimal techniques could not even tackle it properly. So, this work proceeds up by
clustering relevant queries which normally contains the considered, “accessed,” attributes (Hammer and Niamir, 1979), and proceeds down by distributing attributes in clusters so as to all schemes combinations are formed (Navathe, Ceri, Wiederhold and Dou, 1984). The acronym of cluster in this work, however, refers to the fact that acquired query clusters shall be disjoint (Hammer and Niamir, 1979).

The allocation process, on the other hand, has been done in two scenarios. Each scenario (replication/non-replication), is made up of two phases as will be explained in the proposed data allocation model. It is worth mentioning that since this work is set to be for the initial design phase of DDBS, neither retrieval nor update query is given prominence. Meanwhile, the leverage that every type has, as distributed query processed, is explicitly depicted in the evaluation section.

In brief, the main contributions of this paper are summarized as follows:

1. Developing clustering-based vertical fragmentation technique, for initial stage of relational DDBS design, only based on given DDBS queries and their frequencies plus the proposed cost model. In the sense that no need to data statistics, empirical results, midterm predicates or affinities or even attributes affinity matrix;
2. A query refinement process is evolved to assure eliminating overlapping state of obtained fragments (clusters of queries). The resulted schemes of refinement process are taken as inputs to the proposed fragmentation evaluator in purpose of gaining the successful schema. This schema, in its turn, would be considered for two-scenario allocation process in keeping with the proposed allocation cost model;
3. Formulating an objective function for “n-array” fragments in a way that not only reflect the actual transmission, but also guarantee the minimization of communication costs;
4. Presenting convenient site clustering method that network sites gathered into sets of clusters according to their communications costs. This step would naturally contribute in reducing transmission costs as it is clearly drawn in performance evaluation section;
5. Incorporating fragmentation technique using hierarchical clustering, proposing refinement method for obtained fragments and utilizing the existed fragmentation evaluator, network sites clustering, and the proposed data allocation along with data replication techniques into a single effective technique.

The remainder of this paper is organized as follows. Section 2 explores the earlier work which are closely related to this work. Heuristics and architecture of proposed technique are comprehensively presented in section 3. The proposed fragmentation cost model, including objective function and fragmentation cost functions, is given in section 4. Section 5, demonstrates the query and site clustering methodology. In section 6, the proposed allocation and replication model, involving cost functions, are given. Experimental results and performance evaluation are extensively discussed in Section 7. Finally, section 8, presents conclusions and future works directions.

2. RELATED WORK

In DDBS, fragmentation and allocation techniques are essentially developed to lessen the costs of data transmission as distributed query processed. Vertical Fragmentation (VF), in particular, still attracts a great deal of attention due to its huge impact on the enhancement of distributed database management systems DDBMS. In other words, the more accurate VF is, the higher DDBS performance is. As a rule of thumb, when relation precisely partitioned into groups of correlated, “frequently accessed”, attributes and stored together, the DDBS performance is set to be dramatically improved. In (Abuelyaman, 2008), an attribute usage matrix (AUM) and Bond energy algorithm were presented in the sake of introducing an efficient vertical fragment. As a matter of fact, these design methods of
Crowdsourced Knowledge Acquisition: Towards Hybrid-Genre Workflows
[www.igi-global.com/article/crowdsourced-knowledge-acquisition/97651?camid=4v1a](www.igi-global.com/article/crowdsourced-knowledge-acquisition/97651?camid=4v1a)

CRUZAR: An Application of Semantic Matchmaking to e-Tourism
[www.igi-global.com/chapter/cruzar-application-semantic-matchmaking-tourism/38047?camid=4v1a](www.igi-global.com/chapter/cruzar-application-semantic-matchmaking-tourism/38047?camid=4v1a)