Data Mining-based Fragmentation of XML Data Warehouses

Hadj Mahboubi
University of Lyon (ERIC Lyon 2)
5 avenue Pierre Mendès-France
69676 Bron Cedex – France
hadj.mahboubi@eric.univ-lyon2.fr

Jérôme Darmont
University of Lyon (ERIC Lyon 2)
5 avenue Pierre Mendès-France
69676 Bron Cedex – France
jerome.darmont@univ-lyon2.fr

ABSTRACT

With the multiplication of XML data sources, many XML data warehouse models have been proposed to handle data heterogeneity and complexity in a way relational data warehouses fail to achieve. However, XML-native database systems currently suffer from limited performances, both in terms of manageable data volume and response time. Fragmentation helps address both these issues. Derived horizontal fragmentation is typically used in relational data warehouses and can definitely be adapted to the XML context. However, the number of fragments produced by classical algorithms is difficult to control. In this paper, we propose the use of a k-means-based fragmentation approach that allows to master the number of fragments through its k parameter. We experimentally compare its efficiency to classical derived horizontal fragmentation algorithms adapted to XML data warehouses and show its superiority.

Categories and Subject Descriptors
H.2 [Database Management]: Physical Design

General Terms
Performance

1. INTRODUCTION

XML data sources that are pertinent for decision-support become increasingly common with XML becoming a standard for representing complex business data [6]. However, they bear specificities (e.g., heterogeneous number and order of dimensions or complex measures in facts, ragged dimension hierarchies, etc.) that would be intricate to handle in a relational environment. Hence, many efforts toward XML data warehousing have been achieved in the past few years [13, 35, 39], as well as efforts for extending the XQuery language with near On-Line Analytical Processing (OLAP) capabilities, such as advanced grouping and aggregation features [6, 21, 37].

In this context, performance is a critical issue, since XML-native database systems currently suffer from limited performances, both in terms of manageable data volume and response time for complex analytical queries. These issues are typical of data warehouses and can be addressed by fragmentation. Fragmentation consists in splitting a data set into several fragments such that their combination yields the original warehouse without loss nor information addition. Fragmentation can subsequently lead to distribute the target data warehouse, e.g., on a data grid [16]. In the relational context, derived horizontal fragmentation is acknowledged as best-suited to data warehouses, because it takes decision-support query requirements into consideration and avoids computing unnecessary join operations [5]. Several approaches have also been proposed for XML data fragmentation, but they do not take multidimensional architectures (i.e., star-like schemas) into account.

In derived horizontal fragmentation, dimensions’ primary horizontal fragmentation is crucial. In the relational context, two major algorithms address this issue: the predicate construction [32] and the affinity-based [31] strategies. However, both are automatic and the number of fragments is not known in advance, while it is crucial to master it, especially since distributing M fragments over N nodes with $M > N$ can become an issue in itself. Hence, we propose in this paper the use of a k-means-based fragmentation approach that allows to control the number of fragments through its k parameter. Our idea, which adapts a proposal from the object-oriented domain [14] to XML warehouses, is to cluster workload query predicates to produce primary horizontal dimension fragments, with one fragment corresponding to one cluster of predicates. Primary fragmentation is then derived on facts. Queries including given predicates are executed over the corresponding fragments only, instead of the whole warehouse, and thus run faster.

The remainder of this paper is organized as follows. We first discuss existing research related to relational data warehouse, XML data and data mining-based fragmentation in Section 2. Then, we present our k-means-based XML data warehouse fragmentation approach in Section 3. We experimentally compare its efficiency to classical derived horizontal fragmentation algorithms adapted to XML data warehouses and show its superiority in Section 4. Finally, we conclude this paper and present future research directions in Section 5.
2. RELATED WORK

2.1 Fragmentation Definition

There are three fragmentation types in the relational context [23]: vertical fragmentation, horizontal fragmentation and hybrid fragmentation.

Vertical fragmentation splits a relation \( R \) into subrelations that are projections of \( R \) with respect to a subset of attributes. It consists in grouping together attributes that are frequently accessed by queries. Vertical fragments are built by projection. The original relation is reconstructed by joining the fragments.

Horizontal fragmentation divides a relation into subsets of tuples using query predicates. It reduces query processing costs by minimizing the number of irrelevant accessed instances. Horizontal fragments are built by selection. The original relation is reconstructed by fragment union. A variant, derived horizontal fragmentation, consists in partitioning a relation with respect to predicates defined on another relation.

Finally, hybrid fragmentation consists of either horizontal fragments that are subsequently vertically fragmented, or vertical fragments that are subsequently horizontally fragmented.

2.2 Data Warehouse Fragmentation

Many research studies address the issue of fragmenting relational data warehouses either to efficiently process analytical queries or to distribute the warehouse.

To improve ad-hoc query performance, Datta et al. exploit a vertical fragmentation of facts to build the Cuio index [15], while Golfarelli et al. apply the same fragmentation on warehouse views [19]. Munneke et al. propose a fragmentation methodology for a multidimensional database [30]. Fragmentation consists in deriving a global data cube into fragments containing a subset of data. This process is defined by the slice and dice operation. The authors also define another fragmentation strategy, server, that removes one or several dimensions from a hypercube to produce fragments with fewer dimensions than the original data cube. Bellatreche and Boukhalfa apply horizontal fragmentation to a star-schema [5]. Their fragmentation strategy is based on a query workload and exploits a genetic algorithm to select a partitioning schema. This algorithm aims at choosing an optimal fragmentation schema that minimizes query cost. Finally, Wu and Buchmaan recommend to combine horizontal and vertical fragmentation for query optimization [38]. A fact table can be horizontally partitioned with respect to one or more dimensions. It can also be vertically partitioned according to its dimensions, i.e., all the foreign keys to the dimension tables are partitioned as separate tables.

To distribute a data warehouse, Noaman et al. exploit a top-down strategy that uses horizontal fragmentation [32]. The authors propose an algorithm for deriving horizontal fragments from the fact table based on queries that are defined on all dimension tables. Finally, Wehrle et al. propose to distribute and query a warehouse on a computing grid [36]. They use derived horizontal fragmentation to split the data warehouse and build a so-called block of chunks, a data set defining a fragment.

In summary, these proposals generally exploit static derived horizontal fragmentation to reduce irrelevant data access rate and efficiently process join operations across multiple relations [5, 32, 36]. In the literature, the prevalent methods used for derived horizontal fragmentation are the following [23].

- **Predicate construction.** This method fragments a relation by using a complete and minimal set of predicates [32]. Completeness means that two relation instances belonging to the same fragment have the same probability of being accessed by any query. Minimality guarantees that there is no redundancy in predicates.

- **Affinity-based fragmentation.** This method is an adaptation of vertical fragmentation methods to horizontal fragmentation [31]. It is based on the predicate affinity concept [40], where affinity defines query frequency. Specific matrices (predicate usage and affinity matrices) are exploited to cluster selection predicates. A cluster is defined as a selection predicate cycle and forms a dimension graph fragment.

2.3 XML Database Fragmentation

Recently, several fragmentation techniques for XML data have been proposed. They split an XML document into a new set of XML documents. Their main objective is either to improve XML query performance [7, 18, 25] or to distribute or exchange XML data over a network [9, 10].

To fragment XML documents, Ma et al. define a new fragmentation type: split [24, 25], which is inspired from the oriented-object domain. This fragmentation splits XML document elements and assigns a reference to each sub-element. The references are then added to the Document Type Definition (DTD) defining the XML document. The authors extend the DTD and XML-QL languages. Bonfati et al. also propose a fragmentation strategy for XML documents that is driven by structural constraints [7, 8]. This strategy uses both heuristics and statistics. Andrade et al. propose to apply fragmentation to an homogeneous XML collection [2]. They adapt traditional fragmentation techniques to an XML document collection and base their proposal on the Tree Logical Class (TLC) algebra [33]. The authors also evaluate these techniques and show that horizontal fragmentation provides the best performance.

Gertz and Bremer introduce a distribution approach for an XML repository [18]. They propose a fragmentation method and outline an allocation model for distributed XML fragments in a centralized architecture. Gertz and Bremer also define horizontal and vertical fragmentation for an XML document. A fragment is defined with a path expression language, called XF, which is derived from XPath. This fragment is obtained by applying an XF expression on a graph RG representing XML data. Moreover, the authors define exclusion expressions that ensure fragment coherence and disjunction.

Bose and Fegaras use XML fragments for data exchange in a peer-to-peer network (P2P), called XP2P [10]. XML fragments are interrelated and each is uniquely identified by an ID. The authors propose a fragmentation schema, called Tag Structure, to define the structure of data and fragmentation information. Bonfati et al. also define XML fragments for a P2P framework [9]. An XML fragment is obtained and identified by a single path expression, a root-to-node path expression XP, and managed on a specific peer. In addition, the authors associate to each fragment two path expressions:
super fragment and child fragment. These paths ensure the identification of fragments and relationships.

In summary, these proposals adapt classical static fragmentation methods to split XML data. An XML fragment is defined and identified by a path expression [9] or an XML algebra operator [2]. Fragmentation is performed on a single XML document [24, 25] or on an homogeneous XML collection [2]. Note that XML data warehouse fragmentation has not been addressed yet, to the best of our knowledge.

2.4 Data Mining-based Fragmentation

Although data mining has proved useful for selecting physical data structures that enhance performance, such as indexes or materialized views [3, 4], few approaches exploit data mining for fragmentation.

Gorla and Betty exploit association rules (by adapting the Apriori algorithm [1]) for vertical fragmentation approach in relational databases [20].

Darabant and Campan propose the horizontal fragmentation method for object-oriented distributed databases based on k-means clustering we inspire from [14]. This method clusters object instances into fragments by taking all complex relationships between classes into account (aggregation, associations and links induced by complex methods).

Finally, Fiolet and Tourse propose a parallel, progressive clustering algorithm to fragment a database and distribute it over a grid [17]. It is inspired by the CLIQUE sequential clustering algorithm that consists in clustering data by projection.

Though in limited number, these studies clearly demonstrate how data mining can be used for vertical and horizontal fragmentation, through association rule mining and clustering, respectively. They are also static, though.

3. K-MEANS-BASED FRAGMENTATION

Although XML data warehouse architectures from the literature share a lot of concepts (mostly originating from classical data warehousing), they are nonetheless all different. Hence, we proposed a unified, reference XML data warehouse model that synthesizes and enhances existing models [29], and on which we base our fragmentation work. We first recall this model before detailing our fragmentation approach.

3.1 XML Warehouse Reference Model

XML warehousing approaches assume that the warehouse is composed of XML documents that represent both facts and dimensions. All these studies mostly differ in the way dimensions are handled and the number of XML documents that are used to store facts and dimensions. A performance evaluation study of these different representations showed that representing facts in one single XML document and each dimension in one XML document allowed the best performance [12]. Moreover, this representation also allows to model constellation schemas without duplicating dimension information. Several fact documents can indeed share the same dimensions. Hence, we adopt this architecture model. More precisely, our reference data warehouse is composed of the following XML documents:

- a set of dimension.xml documents that each store a given dimension $d$'s member values.

The dw-model.xml document (Figure 1) defines the multidimensional structure of the warehouse. Its root node, DW-model, is composed of two types of nodes: dimension and FactDoc. A dimension node defines one dimension, its possible hierarchical levels (Level elements) and attributes (including their types), as well as the path to the corresponding dimension.xml document. A FactDoc element defines a fact, i.e., its measures, references to the corresponding dimensions, and the path to the corresponding facts.xml document.

![Figure 1: dw-model.xml graph structure](image1)

A facts.xml document stores facts (Figure 2(a)). The document root node, FactDoc, is composed of fact subelements that each instantiate a fact, i.e., measure values and dimension references. These identifier-based references support the fact-to-dimension relationship.

Finally, a dimension.xml document helps instantiate one dimension, including any hierarchical level (Figure 2(b)). Its root node, dimension, is composed of Level nodes. Each one defines a hierarchy level composed of instance nodes that each define the level’s member attribute values. In addition, an instance element contains Roll-up and Drill-Down attributes that define the hierarchical relationship within dimension $d$.

![Figure 2: facts.xml (a) and dimension.xml (b) graph structures](image2)
3.2 Fragmentation Approach

3.2.1 Principle

Since the aim of fragmentation is to optimize query response time, the prevalent fragmentation strategies are workload driven [5, 7, 18, 31, 32]. More precisely, they exploit selection predicates found in workload queries to derive suitable fragments. Our approach also belongs to this family. Its general principle is summarized in Figure 3. It outputs both a fragmentation schema (metadata) and the fragmented XML warehouse. It is subdivided into three steps that are detailed in the following sections:

1. selection predicate extraction from the query workload;
2. predicate clustering with the k-means method;
3. fragment construction with respect to predicate clusters.

![Figure 3: K-means-based fragmentation principle](image)

3.2.2 Selection Predicate Extraction

Selection predicate set $P$ is simply parsed from workload $W$. For example, let $W_S$ be the sample XQuery workload provided in Figure 4 and $P_S$ the corresponding predicate set. $P_S = \{p_1, p_2, p_3, p_4, \ldots \}$, where:

- $p_1 = $y$/attribute[@id="c_nation_key"]/value="15",
- $p_2 = $y$/attribute[@id="c_nation_key"]/value="13",
- $p_3 = $y$/attribute[@id="p_type"]/value="PBC"
- $p_4 = $y$/attribute[@id="d_date_name"]/value="Sat."

For example, $p_2$ and $p_3$ are selection predicates obtained from query $q_2 \in W_S$.

Parsed predicates are then coded in a query-predicate matrix $QP_S$ whose general term $QP_S^j_i$ equals to 1 if predicate $p_i \in P$ appears in query $q_j \in W$, and to 0 otherwise. For example, the $QP_S$ matrix corresponding to $W_S$ and $P_S$ is featured in Table 1.

![Figure 4: XQuery workload snapshot](image)

Table 1: Sample query-predicate matrix $QP_S$

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>$q_2$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$q_{10}$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
</tbody>
</table>

in our case). It attempts to find the centers of natural clusters in source data by minimizing total intra-cluster variance $\sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)^2$, where $C_i, i = 1, \ldots, k$ are the $k$ output clusters and $\mu_i$ is the centroid (mean point) of points $x_j \in C_i$. Let $C$ be the set of all clusters $C_i$.

Usually, having $k$ as an input parameter is viewed as a drawback in clustering. In our case, this turns out to be an advantage, since we want to limit the number of clusters/fragments, typically to the number of nodes the XML data warehouse will be distributed on.

In practice, we used the Weka [22] SimpleKMeans implementation of k-means. SimpleKMeans uses the Euclidean distance to compute distances between points and clusters. It directly inputs matrix $QP$ (actually, the $p_j$ vectors) and $k$, and outputs set of predicate clusters $C$. For example, on matrix $QP_S$ with $k = 2$, SimpleKMeans outputs:

$C_S = \{\{p_1\}, \{p_2, p_3, p_4\}\}$

3.2.3 Predicate Clustering

Our objective is to derive fragments that optimize data access for a given set of queries. Since horizontal fragments are built from predicates, clustering predicates with respect to queries achieves our goal. Intuitively, we ideally seek to build rectangles (clusters) of 1s in matrix $QP$. We chose the widely-used k-means algorithm [26] for clustering. This algorithm inputs vectors of object attributes (columns of $QP$)
Figure 5: Fragment construction substeps

Figure 6: frag-schema.xml graph structure

fragments.xq script) that, applied to the XML data warehouse (i.e., the whole set of facts1.xml and dimension1.xml documents), produces the actual fragments, which we store in a set of facts2.xml and dimension2.xml documents, $i = 1, ..., k + 1$. As fragments, these documents indeed bear the same schema than the original warehouse. The $(k + 1)^{th}$ fragment is based on an additional predicate, denoted ELSE, which is the negation of the conjunction of all predicates in $P$ and is necessary to ensure fragmentation completeness (Section 2.2). In our running example, $ELSE = \neg (p_1 \land p_2 \land p_3 \land p_4)$.

Figure 8 provides an excerpt from the fragments2.xq script that helps build fragment $f_2$ from Figure 7. Dimension fragments are generated first, one by one, through selections exploiting the predicate(s) associated to the current dimension (three first queries from Figure 8). Then, fragmentation is derived on facts by joining the original fact document to the newly-created dimension fragments (last query).

4. EXPERIMENTS

Since derived horizontal fragmentation is a NP-hard problem [11] solved by heuristics, we choose to validate our proposal experimentally.

4.1 Experimental conditions

We use XWeB (XML Data Warehouse Benchmark) [27] as a test platform. XWeB is based on the reference model defined in Section 3.1, and proposes a test XML data warehouse and its associated XQuery decision-support workload.

Figure 7: Sample frag-schema.xml document


Figure 8: Excerpt from sample fragments2.xq script

XWeB's warehouse consists of sale facts characterized by the amount (of purchased products) and quantity (of purchased products) measures. These facts are stored in the facts2.xml document and are described by four dimensions: Customer, Supplier, Date and Part stored in the dimension1.xml document.

XWeB's workload is composed of queries that exploit the warehouse through join and selection operations. We extend this workload by adding queries and selection predicates in order to obtain a significant fragmentation. Due to space constraints, our workload is only available on-line.

We ran our tests on a Pentium 2 GHz PC with 1 GB of main memory and an IDE hard drive under Windows XP. We use the X-Hive XML native DBMS² to store and query the warehouse. Our code is written in Java and connects

1 http://eric.univ-lyon2.fr/~hmahloufi/Workload/workload.xq
2 http://www.x-hive.com/products/dh/
to X-Hive and Weka through their respective Application Programming Interfaces (APIs). It is available on demand.

### 4.2 Fragmentation Strategy Comparison

In this first series of experiments, we aim at comparing our k-means-based horizontal fragmentation approach (denoted KM) to the classical derived horizontal fragmentation techniques, namely by predicate construction (PC) and affinity-based (AB) primary fragmentation (Section 2.2), which we adapted to XML data warehouses [28]. We also record performance when no fragmentation is applied (NF), for reference.

#### 4.2.1 Query Response Time

This experiment measures workload execution time with the three fragmentation strategies we adopted. For KM, we arbitrarily fixed \( k = 8 \), which could correspond to a computer cluster’s size. The fragments we achieve are stored in distinct collections to simulate data distribution. Each collection can indeed be considered to be stored on a distinct node and can be identified, targeted and queried separately.

To measure query execution time over a fragmented warehouse, we first identify the required fragments with the \texttt{frag-schema.xml} document. Then, we execute the query over each fragment and save execution time. To simulate parallel execution, we only consider maximum execution time.

Figure 9 plots workload response time with respect to data warehouse size (expressed in number of facts). It clearly shows that fragmentation significantly improves response time, and that KM fragmentation performs better than PC and AB fragmentation when the warehouse scales up. Workload execution time is indeed, on an average, 86.5% faster with KM fragmentation than with NF, and 36.7% faster with KM than with than AB. We believe our approach performs better than classical derived horizontal fragmentation techniques because these latter produce many more fragments (159 with PC and 119 with AB vs. 9 with KM). Hence, at workload execution time, queries must access many fragments (up to 50 from our observations), which multiplies query distribution and result reconstruction costs. The number of accessed fragments is much lower with KM (typically 2 fragments in our experiments).

#### 4.2.2 Fragmentation Overhead

We also compare the PC, AB and KM \((k = 8)\) fragmentation strategies in terms of overhead (i.e., fragmentation algorithm execution time). When assessing performance, it is indeed necessary to find a fair trade-off between gain and overhead. Table 3 summarizes the results we obtain for an arbitrarily fixed data warehouse size of 3,000 facts. It shows that KM clearly outperforms AB and PC, which is in line with these algorithms' complexities: \(O(|F|)\), \(O(|F|^2)\) and \(O(2^{|P|})\), respectively. While AB and PC would have to run off-line, KM could on the other hand be envisaged to run on-line.

#### 4.3 Influence of Number of Clusters

In this experiment, we fixed data warehouse size (to 4,000 and 5,000 facts, respectively) and varied KM parameter \( k \) to observe its influence on workload response time. Figure 10 confirms that performance improves quickly when fragmentation is applied, but tends to degrade when the number of fragments increases, as we explained in Section 4.2.1. Furthermore, it hints that an optimal number of clusters for our test data warehouse and workload lies between 4 and 6, making us conclude that over-fragmentation must be detected and avoided. Note that, on Figure 10, \( k = 1 \) corresponds to the NF experiment (this one fragment is the original warehouse).

### 5. CONCLUSION

In this paper, we have introduced an approach for fragmenting XML data warehouses that is based on data mining, and more precisely clustering and the k-means algorithm. Classical derived horizontal fragmentation strategies run automatically and output an unpredictable number of fragments, which is nonetheless crucial to keep under control. By contrast, our approach allows to fully master the
number of fragments through the k-means k parameter.

To validate our proposal, we have compared our fragmentation strategy to XML adaptations of the two prevalent fragmentation methods for relational data warehouses. Our experimental results show that our approach, by producing a lower number of fragments, outperforms both the others in terms of performance gain and overhead.

Now that we have efficiently fragmented an XML data warehouse, our more direct perspective is to distribute it on a data grid. This raises several issues that include processing a global query into subqueries to be sent to the right nodes in the grid, and reconstructing a global result from subquery results. Properly indexing the distributed warehouse to guarantee good performance shall also be very important.

Finally, in a continuous effort to minimize the data warehouse administration function and aim at autoadministration systems [3, 4], we plan to make our data mining based-fragmentation strategy dynamic. The idea is to perform incremental fragmentation when the warehouse is refreshed. This could be achieved with the help of an incremental variant of the k-means algorithm [34].

6. ACKNOWLEDGMENTS

The authors would like to thank Housssem Aissa, Anouar Benzakour, Kevin du Repaire and Hamza El Kartite, who participated in coding our approach in Java.

7. REFERENCES


