

A Relevance-Extended Multi-dimensional Model for a Data Warehouse Contextualized with Documents

Juan Manuel Pérez, Rafael Berlanga,
María José Aramburu
Universitat Jaume I
{martinej,berlanga,aramburu}@uji.es

Torben Bach Pedersen
Aalborg University
tbp@cs.aau.dk

ABSTRACT

Current data warehouse and OLAP technologies can be applied to analyze the structured data that companies store in their databases. The circumstances that describe the context associated with these data can be found in other internal and external sources of documents. In this paper we propose to combine the traditional corporate data warehouse with a document warehouse, resulting in a contextualized warehouse. Thus, contextualized warehouses keep a historical record of the facts and their contexts as described by the documents. In this framework, the user selects an analysis context which is represented as a novel type of OLAP cube, here called *R-cube*. *R-cubes* are characterized by two special dimensions, namely: the relevance and the context dimensions. The first dimension measures the relevance of each fact in the selected analysis context, whereas the second one relates each fact with the documents that explain their circumstances. In this work we extend an existing multi-dimensional data model and algebra for representing the *R-cubes*.

Categories and Subject Descriptors

H.4.2 [Information Systems Applications]: Types of Systems—*decision support*

General Terms

Design

Keywords

OLAP, text-rich XML documents

1. INTRODUCTION

Current data warehouse [8] and OLAP [3] technologies can be efficiently applied to analyze the huge amounts of structured data that companies produce. These organizations also produce many documents and use the Web as

their largest source of external information. Examples of internal and external sources of information include the following: purchase-trends and market-research reports, demographic and credit reports, popular business journals, industry newsletters, technology reports, etc. Although these documents cannot be analyzed by current OLAP technologies mainly because they are unstructured and contain a large amount of text, they include highly valuable information that should also be exploited by companies. The current trend is to find these documents available in XML-like formats [21].

Our proposal is to build XML document warehouses that can be used by companies to store unstructured information coming from their internal and external sources. In [16] we outlined a multi-dimensional implementation of a document model [15] for the analysis of the information stored in a warehouse of text-rich XML documents. In this paper we present an architecture for the integration of a corporate warehouse of structured data with a warehouse of unstructured documents. We call the resulting warehouse a *contextualized warehouse*. Thus, a contextualized warehouse is a new kind of decision support system that allows users to obtain strategic information by combining all their sources of structured data and unstructured documents, and by analyzing the integrated data under different contexts. For example, if we have a document warehouse with business news articles, we can analyze the evolution of the sales measures stored in our corporate warehouse in the context of a period of crisis as described by the relevant news. Thus, we could detect which products have been more affected. The same set of facts could be less revealing under a different context (e.g. regions in economical development). Furthermore, for those facts that are not available in the corporate warehouse, some kind of alternative approximate information about them could be extracted from past economical reports (e.g. aggregated measures of historical export-import rates of some countries), and then included in the contextualized warehouse. We note that some important characteristics make different typical OLAP facts from factual information extracted from documents: the extracted facts may be incomplete (since not all the dimensions may be quoted in the documents contents) and/or imprecise (if the dimension values found belong to non-base levels).

The applications described above require both the availability of a document warehouse and its cooperation with the corporate data warehouse. The circumstances of the original facts are understood by analyzing their contexts, that is, the information available in the documents related

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DOLAP'05, November 4–5, 2005, Bremen, Germany.
Copyright 2005 ACM 1-59593-162-7/05/0011 ...\$5.00.

with the facts. In this paper, a context is defined as a *set of textual fragments that can provide analysts with strategic information important for decision-making tasks*. Contexts are thus unstructured, and cannot be managed by the well-structured corporate warehouse. Since the document warehouse may contain documents about many different topics, we apply well-known Information Retrieval (IR) [1] techniques to select the context of analysis from the document warehouse. Thus, in order to build a contextualized OLAP cube, the analyst will specify the context under analysis by supplying a sequence of keywords. Each fact in the resulting cube will have a numerical value representing its relevance with respect to the specified context, thereby its name, *R-cube* (Relevance cube). Moreover, each fact in the *R-cube* will be related to its context (i.e. the set of the relevant documents that describe the context of the fact). In this paper we extend an existing multi-dimensional data model to represent these two new dimensions (relevance and context), and we study how the traditional OLAP operations affect them.

The relevance and context dimensions provide us further information about facts that can be very useful for analysis tasks. From the user point of view, the relevance dimension can be used to explore the most relevant portions of an *R-cube*. For example, it can be used to identify the period of a political crisis, or the regions under economical development. The usefulness of the context dimension is twofold. First, it can be used in the selection operations to restrict the analysis to the facts described in a given subset of the documents (e.g. the most relevant documents). And second, the user will be able to gain insight into the circumstances of a fact by retrieving its related documents. The graphs, charts, and other binary files possibly linked in the documents would also be presented to the user, easing the understanding of the analysis context.

The main contributions of this paper are both, (1) an architecture for the integration of a traditional structured corporate warehouse with a document warehouse, resulting in a *contextualized warehouse*; and (2) the formal multi-dimensional data model and the algebra unary operations to manage *R-cubes*.

The rest of the paper is organized as follows. Section 2 discusses the related work. In Section 3 we present the architecture of a contextualized warehouse. Section 4 shows how the analysis cubes (*R-cubes*) are built. The multidimensional data model for *R-cubes* is presented in Section 5. In Section 6 we propose an algebra for *R-cubes*. Finally, Section 7 addresses conclusions and future work.

2. RELATED WORK

In [7] the importance of external contextual information to understand the results of historical analysis was emphasized. Some works like [21] are focused on the construction of repositories of XML documents gathered from the Internet, but they do not provide on-line analysis tools.

In [13], OLAP operations are extended to involve dimension and/or measures contained in external XML data. From a different point of view, [2] propose to extend XQuery [20] with grouping constructs to evaluate OLAP-style aggregation queries on XML documents. However, the cited papers only deal with highly structured XML data (e.g. on-line XML products pricing lists), since the measures and dimensions should be selected by using XPath expressions [20].

These approaches are not suitable for analyzing text-rich XML documents where the measure and dimension values are found in the documents textual contents.

A recent paper [12] provides mechanisms to perform special text aggregations on the contents of XML documents, e.g., getting the number of words of a document section, its most frequent keywords, a summary, etc. Although these text-mining operations are very useful to explore an XML document collection, these techniques cannot be applied to evaluate OLAP operations on the factual information described by the textual contents of the documents. Our approach differs from [12] in the sense that we do not analyze the documents textual contents themselves, we extract the dimension values from the documents contents and relate the documents with those corporate facts characterized by the same dimension values. Afterwards, we analyze the corporate data by using the documents as their context.

Information Retrieval (IR) [1] is playing an important role on the Web, since it has enabled the development of useful discovery tools (e.g. web search engines) and digital library services. These applications deal with huge amounts of text-rich documents and have successfully applied IR techniques to query this type of repositories. In an IR system the users describe their information needs by supplying a sequence of keywords. The result is a set of documents ranked by relevance. The relevance is a numerical value which measures how well the document fits the user information needs. Traditional IR models (e.g. the vector space model [18]) calculate this relevance value by considering the local and global frequency (tf-idf) of the query keywords in the document and the collection, respectively. Intuitively, a document will be relevant to the query if the specified keywords appear frequently in its textual contents and they are not frequent in the collection. Newer proposals in the field of IR include language modeling [17] and relevance modeling [9] techniques. The works on language modeling consider each document as a language model. Thus, documents are ranked according to the probability of obtaining the query keywords when randomly sampling from the respective language model. An extension of the language modeling approach is relevance modeling [9] which estimates the probability of observing a query keyword in the set of documents relevant to a query. The language and relevance modeling approaches still internally apply the keyword frequency to estimate probabilities, and they have been shown to outperform baseline tf-idf models in many cases [17, 9]. As discussed below, our approach relies on relevance modeling techniques. Unlike traditional IR models, language and relevance models provide a formal background based on probability theory which is suitable to be included in the formalization of OLAP operations.

In [15] we presented a document model for text-rich XML documents where information extraction techniques [10, 5, 6] are applied to identify the facts described by the textual contents of the documents. Particularly, [10] and [5] were proven to work to extract time and location references. Given an IR condition, in [15] we showed how relevance modeling techniques [9] can be applied to estimate the relevance of a fact by the probability of observing this fact in a document which contains the keywords stated in the IR condition. By following this research line, in a short paper [16] we outlined a multi-dimensional implementation of the relevance model discussed in [15]. The focus of the present paper is contextualized warehouses. Here we relate

the facts of a traditional corporate warehouse with the documents that describe their circumstances. IR conditions are used for establishing an analysis context, and the relevance model of [15] is applied to calculate the relevance of the facts in the resulting analysis cube (*R-cube*). The relevance value of a fact measures how well the fact fits into the selected context. At any moment, it is possible to retrieve the documents related to a fact, and to gain insight into the explanation of its circumstances. The architecture of a contextualized warehouse, and the extended multi-dimensional data model and algebra for *R-cubes* proposed in the current paper are novel.

In [11] a probabilistic multi-dimensional model was presented, but it does not clearly discuss how the probabilities of the facts are re-calculated after an aggregation. Moreover, the selection operation does not modify the probabilities of the facts in the resulting cube. We consider that the conditions established in a selection operation are also a restriction on the context under analysis and therefore the relevance of facts must be updated accordingly.

3. CONTEXTUALIZED WAREHOUSES

Figure 1 shows the proposed architecture for the contextualized warehouse. Its main components are a traditional structured corporate warehouse [8], a document warehouse able to evaluate IR conditions [15], and a fact extractor module. Building a contextualized warehouse mainly means relating each fact of the corporate warehouse with its context. For this purpose, the fact extractor tool uses the dimensions defined in the corporate warehouse to detect and build the facts described in the documents. Next, we describe this process in detail by means of an example.

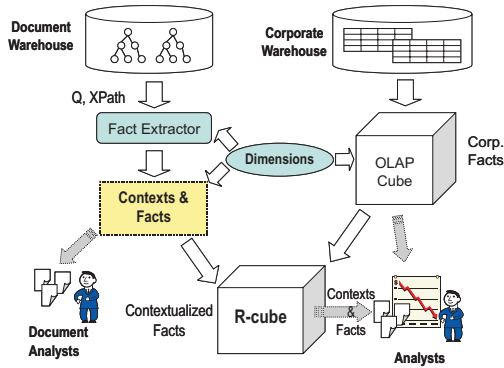


Figure 1: Contextualized warehouse architecture

Let us consider the corporate warehouse of an international provider of vegetable oil by-products. The main products of this company include: *fo1*, *fo2* (used as preservatives in the *food* sector), and *he1* and *he2* (used in the elaboration of *healthcare* products). The company keeps in its corporate warehouse a historical record of its sales, the quantity sold (*Quantity* measure) and its cost (*Amount* measure), per product and customer. Thus, the dimensions of the corporate warehouse are *Time*, *Products* and *Customers*. The *Products* are classified into *Sectors* (*food* and *healthcare*). Finally, *Customers* are organized into *Countries* and *Regions* (e.g. Southeast Asia, Central America, etc.).

Our example company also maintains a document warehouse of business newspapers gathered from the Internet in XML format. Figure 2 shows a fragment of an example document of this warehouse that depicts a context for the sales of food sector products to customers of the Southeast Asian region, made during the second half of 1998. Notice that contexts are very useful for analysis tasks, since they can give us detailed information about the facts of the corporate warehouse. For example, the document in Figure 2 would help to understand a sales drop.

```
<article date='Dec.1,1998'>
<paragraph>
The financial crisis in Southeast Asian countries,
has mainly affected companies in the food market
sector. Particularly, Chicken SPC Inc. has reduced
total exports to $1.3 million during this half of the
year from $10.1 million in 1997.
</paragraph> ...
</article>
```

Figure 2: Example fragment of a business journal

By applying information extraction techniques [10, 5, 6], and considering the predefined analysis dimensions of the example corporate warehouse, the dimension values *Southeast Asia*, *food*, and *1998/2nd half* can be identified in the document fragment. The fact extractor tool builds all the valid facts with them, in this case, (*Products.Sector = food*, *Customers.Region = Southeast Asia*, *Time.Half_year = 1998/2nd half*). As it can be noticed, some of these dimension values are not *precise* enough and belong to non-base dimension categories, for example, the *SoutheastAsia* dimension value belongs to the category *Region* of the *Customers* dimension. We may also find documents where some dimensions are not mentioned, resulting in *incomplete* facts. For each fact, the fact extraction tool also keeps the number of times that its dimension values occur in the document fragment (i.e. the fact dimension values frequency). This frequency determines the importance of the fact in the document, and later will be used to estimate the relevance of a fact in a given context.

Let us now consider the second sentence of the example document of Figure 2. It depicts two facts: (*Company = Chicken SPC*, *Time.Year = 1997*, *Export = \$10,100,000*), (*Company = Chicken SPC*, *Time.Half_year = 1998/2nd half*, *Export = \$1,300,000*). Chicken SPC Inc. could be a potential customer or competitor of our example oil provider company. In this way, the document warehouse also provides highly valuable strategic information about some facts that are not available in the corporate warehouse nor in external databases. We note that sometimes it is relatively easy to obtain these facts, for example, when they are presented as tables in the documents. However, most times documents contain already aggregated measure values (total exports in the facts of the previous example). The main problem here is to automatically infer the implicit aggregation function that was applied (i.e. average, sum, etc.) Alternatively, the system could ask the user to guess the aggregation function by showing him/her the document contents. In this work we mainly focus on the fact *dimensions*, leaving for future work the management of measures extracted from texts. Notice that documents extracted measure values are not essential to construct a contextualized warehouse, since the dimension values found in a document are sufficient to relate it

| F | Products.ProductId | Customers.Country | Time.Month | Amount | R | Ctxt |
|-------|--------------------|-------------------|------------|-------------|------|----------------------------|
| f_1 | $fo1$ | <i>Cuba</i> | 1998/03 | 4,300,000\$ | 0.05 | $d_3^{0.005}, d_7^{0.005}$ |
| f_2 | $fo2$ | <i>Japan</i> | 1998/02 | 3,200,000\$ | 0.1 | $d_5^{0.02}$ |
| f_3 | $fo2$ | <i>Korea</i> | 1998/05 | 900,000\$ | 0.2 | $d_4^{0.04}$ |
| f_4 | $fo1$ | <i>Japan</i> | 1998/10 | 300,000\$ | 0.4 | $d_1^{0.04}, d_2^{0.08}$ |
| f_5 | $fo2$ | <i>Korea</i> | 1998/11 | 400,000\$ | 0.25 | $d_2^{0.08}, d_6^{0.01}$ |

Table 1: Example R -cube. Each row represents a fact. The R and the $Ctxt$ columns (dimensions) depict the relevance value and the context of the facts, respectively. Each d_i^r denotes a document fragment of the collection whose relevance with respect to Q is r .

with the corporate facts that are characterized by these dimension values.

4. BUILDING R -CUBES

In this section we study how the analysis cubes are materialized from the contextualized warehouse. We call them R -cubes since they include two special system-maintained dimensions, namely: the *relevance* and the *context* dimensions. From these cubes, users can study the contextualized facts.

In order to create an R -cube the analyst must supply a query of the form $(Q, XPath, MDX)$, which states the following restrictions: Q is an Information Retrieval (IR) condition, consisting of a sequence of keywords that specifies the context under analysis; $XPath$ is a path expression [20] that establishes the document sections under study; finally, MDX are conditions over the dimensions and measures of the analysis [19]. Here, our purpose is not to define a new query language, but to identify the type of conditions needed to build an analysis cube in a contextualized warehouse.

The query process takes place as follows: (1) First, the IR condition Q and the path expression $XPath$ are evaluated in the documents warehouse. The result is a set of documents fragments satisfying $XPath$ and Q , along with their relevance with respect to Q . (2) Second, the fact extractor component parses the documents fragments obtained in step (1) and returns the set of facts described by each document fragment, along with their frequency. Notice that we do not parse entire documents, but those document fragments selected by the $XPath$ expression. (3) Next, or in parallel to steps (1) and (2), the MDX conditions are evaluated on the corporate warehouse. (4) Then, we assign each document to those facts of the corporate database whose dimension values can be “rolled-up” or “drilled-down” to some (possibly imprecise or incomplete) fact described by this document. (5) Finally, we calculate the relevance of each fact, resulting in an R -cube.

By following with the running example, let us consider the analysis of the sales of food products under the context of a financial crisis reported in the business articles of the document warehouse. Thus, given $Q = \text{“financial, crisis”}$, $XPath = \text{“/db/business/article/paragraph”}$ and $MDX = (Products.[food], Customers.Country, Time.[1998].Month, SUM(Measures.Amount) > 0)$ as query conditions, the contextualized warehouse will return the R -cube presented in Table 1. That is, the set of facts of the corporate warehouse that satisfy the stated MDX conditions, along with their relevance values with respect to the IR condition (relevance dimension, depicted as R), and the set of text fragments where each fact occurs (context dimension, represented by $Ctxt$).

As Table 1 shows, the relevance (R dimension) is a numeric value that measures the importance of each fact in the context established by the initial query conditions. The most relevant facts of our example R -cube involve the sales made to Japanese and Korean customers during the months of October and November 1998. Notice that we could obtain the details described in the relevant documents by performing a drill-through operation on the context dimension [19]. By studying these documents we can discover that the Southeast Asian financial crisis reported by the document of Figure 2, is a valid explanation of the sales drop. Each document d_i of the context dimension also has associated a relevance value (represented by the r superscript in d_i^r) which measures how well this document describes the selected analysis context.

A detailed discussion on the calculus of the relevance of the facts can be found in [15], a brief explanation of the involved formulas is included in Appendix A. Intuitively, a fact will be relevant for the selected context if the fact is found in a document which is also relevant for this context (e.g. if the keywords of the IR condition Q occur frequently in the document). In [15] we applied relevance modeling techniques [9] to estimate the relevance of a fact by means of the probability of observing this fact in the set of documents relevant to the IR condition. The probability of finding a fact in a document is determined by the frequency of the dimension values of this fact in the textual contents of the document.

An interesting property of this approach is that the sum of the relevance values of all the facts in an R -cube is equal to one. However, notice that not all the document collections are suitable for all the analysis (e.g. an analysis on financial crisis with a document collection about products manufacturing processes) and the sum of the facts relevance values will be kept equal to one. Later in the paper, we will denote by *Quality* the factor that normalizes the relevance values of the facts. This factor indicates the quality of the R -cube for the selected context, since it measures the overall relevance of the documents that satisfy the IR condition Q .

Unlike OLAP-XML federations like those proposed in [13], R -cubes are materialized once, when the query is fetched to the contextualized warehouse, and will be incrementally updated when new relevant documents and data satisfying the original query are added to their respective warehouses. The main advantage of this approach is that pre-aggregations can be performed over R -cubes, thus allowing fast analysis operations over them.

The rest of the paper will focus on the formal definition of R -cube, and the provision of a suitable algebra for OLAP operations.

5. A MULTIDIMENSIONAL DATA MODEL FOR R -CUBES

In this section we define a data model for the R -cubes. We extend an existing multidimensional model [14] with two new special dimensions to represent both, the relevance of the facts and their context. For each component of the extended data model, we show its definition and give some examples.

5.1 Dimensions

A *dimension* D is a two-tuple $D = (C_D, \sqsubseteq_D)$, where $C_D = \{C_j\}$ is a set of categories C_j .

Example 1. In [14] everything that characterizes a fact is considered to be a dimension, even those attributes modeled as measures in other approaches. Figure 3 shows the dimensions for the running example.

Each category $C_j = \{e\}$ is a set of dimension values. \sqsubseteq_D is a partial order on $\cup_j C_j$ (the union of all dimension values in the individual categories). Given two values $e_1, e_2 \in \cup_j C_j$, then $e_1 \sqsubseteq_D e_2$ if e_1 is logically contained in e_2 . The intuition is that each category represents the values of a particular granularity. We will write $e \in D$, e is a dimensional value of D , if $e \in \cup_j C_j$.

There are two special categories present in all dimensions: $\top_D, \perp_D \in C_D$ (the top and bottom categories). The category \perp_D has the values with the finest granularity. These values do not logically contain other category values and are logically contained by the values of other coarser categories. The category $\top_D = \{\top\}$ represents the coarsest granularity. For all $e \in D, e \sqsubseteq_D \top$.

The partial order \sqsubseteq_D can be generalized to work on categories as follows: given $C_1, C_2 \in C_D$, then $C_1 \sqsubseteq_D C_2$ if $\exists e_1 \in C_1, e_2 \in C_2, e_1 \sqsubseteq_D e_2$. We will write \sqsubseteq instead of \sqsubseteq_D when it is clear that \sqsubseteq is the partial order of the dimension D .

Example 2. The *Customers* dimension has the categories $\perp_{Customers} = Country \sqsubseteq Region \sqsubseteq \top_{Customers}$, with the dimension values $Country = \{Japan, Korea, Cuba, \dots\}$ and $Region = \{Southeast Asia, Central America, \dots\}$. The partial order on category values is: $Japan \sqsubseteq Southeast Asia \sqsubseteq \top, Korea \sqsubseteq Southeast Asia, Cuba \sqsubseteq Central America \sqsubseteq \top$, etc.

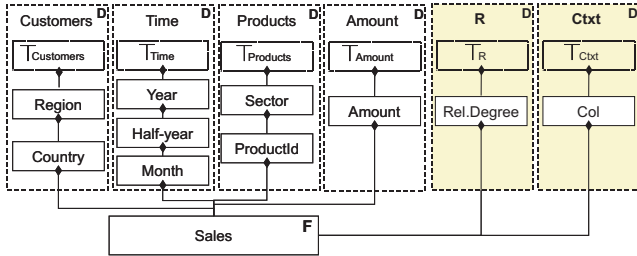


Figure 3: Dimensions of the example case of study. R and $Ctxt$ are the relevance and context dimensions of the R -cube.

The Relevance Dimension. The *relevance* dimension depicts the importance of each fact of the R -cube in the selected context (i.e. the IR condition Q). Therefore, it can

be used to identify the portions of an R -cube that are more interesting for the context of analysis.

Different approaches can be followed to state the *relevance* dimension R . The simplest one is to define it just with the bottom and top categories: $\perp_R = Relevance \sqsubseteq_R \top_R$. Since we model the relevance as a probability value, the values of the *Relevance* category are real numbers in the interval $[0,1]$. Like in [11], we propose to introduce an intermediate category to allow users to study relevance values from a higher qualitative abstraction level. In this new category the relevance values will be classified into groups (*Relevance Degrees*) like *irrelevant*, *relevant* or *very relevant*.

However, the relevance values are normalized to have their sum equal to one. Thus, a relevance index of 0.02 may be *irrelevant* if the rest of relevance values are significantly greater, or *relevant* if the maximum value of relevance obtained was, for example, 0.03.

Thus, we need to define a dynamic partial order \sqsubseteq_R^γ to map the values r of the base *Relevance* category to values of the *Relevance Degree* category depending on the value of r/γ . In this way, we will use γ as a normalization factor. Note that γ should measure the global relevance of a particular result. Typical measures are $\gamma = MAX(r)$, $\gamma = AVG(r)$ or $\gamma = Quality$.

Definition 1. The *relevance dimension* is a two-tuple $R = (C_R, \sqsubseteq_R^\gamma)$ where: $C_R = \{Relevance, Relevance Degree, \top_R\}$ is the set of categories; $Relevance = [0,1]$ is the base category \perp_R ; $Relevance Degree = \wp([a,b])$ is a partition on the interval of Real numbers $[a,b]$; and \sqsubseteq_R^γ is the partial order $r \sqsubseteq_R^\gamma rd$, if $r \in Relevance, rd \in Relevance Degree$ and $r/\gamma \in rd$.

Example 3. Let us consider $\gamma = MAX(r)$ (the maximum value of relevance obtained in the R -cube), and five different degrees of relevance, $Relevance Degree = \{very\ irrelevant = [0, 0.25], irrelevant = [0.25, 0.45], neutral = [0.45, 0.55], relevant = [0.55, 0.75], very\ relevant = [0.75, 1]\}$, a partition of $[0,1]$. As Table 1 shows, $MAX(r) = 0.4$, then $0.05 \sqsubseteq_R^{0.4} very\ irrelevant, 0.1 \sqsubseteq_R^{0.4} irrelevant, 0.2 \sqsubseteq_R^{0.4} neutral, 0.4 \sqsubseteq_R^{0.4} very\ relevant$ and $0.25 \sqsubseteq_R^{0.4} relevant$.

The Context Dimension. The context of the facts is detailed by the documents of the warehouse. We represent these documents in the *context* dimension, along with their relevance to the IR condition Q .

Definition 2. The *context dimension* is a two-tuple $Ctxt = (C_{Ctxt}, \sqsubseteq_{Ctxt})$, where $C_{Ctxt} = \{Col, \top_{Ctxt}\}$ is the set of categories. The category $\perp_{Ctxt} = Col = \{d^r\}$ is the set of the documents d of the warehouse, the subscript r denotes the relevance of the document to the context of analysis (the IR condition Q).

Example 4. In our example, d_1, \dots, d_7 are documents of the warehouse which describe the context of the facts presented in the R -cube. The relevance value with respect to the established IR condition is 0.04 for d_1 , 0.08 for d_2 , 0.005 for d_3 , 0.04 for d_4 , 0.02 for d_5 , 0.01 for d_6 and 0.005 d_7 . Then, $\{d_1^{0.04}, d_2^{0.08}, d_3^{0.005}, d_4^{0.04}, d_5^{0.02}, d_6^{0.01}, d_7^{0.005}\} \subset Ctxt$.

The context dimension as defined in Definition 2 is plain, i.e., it has no hierarchies. It is possible to define a hierarchy for the context dimension by considering the hierarchical structure of the XML documents, and by classifying

the different documents fragments into the category which represents their element type (i.e. tag name). This hierarchy would allow the user to navigate on the structure of the XML documents in the OLAP operations. However, for simplicity reasons we do not include such a hierarchy in the paper.

5.2 Fact-Dimension Relations

The fact-dimension relations link facts with dimension values. By following [14], given a set of facts $F = \{f\}$ and a dimension D , the *fact-dimension relation* between F and D is the set $FD = \{(f, e)\}$, where $f \in F$ and $e \in D$.

A fact f is *characterized* by the dimension value e , written $f \sim_D e$, if $\exists e' \in D, (f, e') \in FD \wedge e' \sqsubseteq_D e$. In order to avoid missing values it is required that $\forall f \in F, \exists e \in D, (f, e) \in FD$. If the dimension value that characterizes a fact is not known, the pair (f, \top) is added to FD .

Example 5. In the example of Table 1 we have the facts $F = \{f_1, f_2, f_3, f_4, f_5\}$. $FD_{Customers}$ is the fact-dimension relation that links each fact with its value in the dimension *Customers*. Thus, $FD_{Customers} = \{(f_1, Cuba), (f_2, Japan), (f_3, Korea), (f_4, Japan), (f_5, Korea)\}$, and for example $f_3 \sim_{Customers} Southeast\ Asia$.

Example 6. The fact-dimension relation FD_{Amount} links each fact with its value in the dimension *Amount*, $FD_{Amount} = \{(f_1, 4, 300, 000\$), (f_2, 3, 200, 000\$), (f_3, 900, 000\$), (f_4, 300, 000\$), (f_5, 400, 000\$)\}$.

The Relevance Fact-Dimension Relation. The relevance fact-dimension relation links each fact with its relevance value.

Definition 3. The *relevance fact-dimension relation* is the set $FR = \{(f, r)\}$ where $f \in F$ is a fact and $r \in R$ its relevance. We require each fact to have a unique relevance value, $\forall f \in F, \exists! r \in R, (f, r) \in FR$. The sum of the relevance values of all the facts in F is equal to one, $\sum_{(f,r) \in FR} r = 1$.

Let $rd \in RelevanceDegree$ and γ , we will write $f \sim_R^\gamma rd$, meaning that the relevance degree of the fact f is rd when global relevance measure γ is applied, if $\exists r \in R, (f, r) \in FR$ and $r \sqsubseteq_R^\gamma rd$.

Example 7. For the running example we have $FR = \{(f_1, 0.05), (f_2, 0.1), (f_3, 0.2), (f_4, 0.4), (f_5, 0.25)\}$, and by taking $\gamma = MAX(r) = 0.4$, $f_1 \sim_R^{0.4} very\ irrelevant$, $f_2 \sim_R^{0.4} irrelevant$, $f_3 \sim_R^{0.4} neutral$, $f_4 \sim_R^{0.4} very\ relevant$ and $f_5 \sim_R^{0.4} relevant$. That is, f_5 is *relevant*, but f_2 may be *irrelevant* for the selected context.

The Context Fact-Dimension Relation. The context fact-dimension relation links each fact with the documents that describe its context.

Definition 4. We define the *context fact-dimension relation* as the set $FCtxt = \{(f, d'')\}$ where $f \in F$ is a fact described by the document d , also written $f \sim_{Ctx} d$, and r is the relevance value of d , $d'' \in Ctxt$.

Example 8. In the example, $FCtxt = \{(f_1, d_3^{0.005}), (f_1, d_7^{0.005}), (f_2, d_5^{0.02}), (f_3, d_4^{0.04}), (f_4, d_1^{0.04}), (f_1, d_2^{0.08}), (f_5, d_2^{0.08}), (f_5, d_6^{0.01})\}$. The documents d_1, d_2 depict the context of the fact f_4 , then $f_4 \sim_{Ctx} d_1$ and $f_4 \sim_{Ctx} d_2$.

5.3 R-cubes: Relevance-Extended Multidimensional Objects

We extend the definition of *multidimensional object* [14] to include the relevance and context dimensions discussed before.

Definition 5. A *relevance-extended multidimensional object* or *R-cube* is a four-tuple $RM = (F, D, FD, Q)$, where: $F = \{f\}$ is a set of facts; $D = \{D_i, i = 1, \dots, n\} \cup \{R, Ctxt\}$ is a set of dimensions, $R, Ctxt \in D$ are the relevance and context dimensions previously defined; $FD = \{FD_i, i = 1, \dots, n\} \cup \{FR, FCtxt\}$ is a set of fact-dimension relations, one for each dimension $D_i \in D$; $FR, FCtxt \in FD$ are the relevance and context fact-dimension relations discussed above; and Q is an IR condition. In the model, we represent the relevance of each fact to the context established by the IR condition Q .

We measure the analysis quality of an *R-cube* RM for the selected context by $Quality = \sum_{(f,d'') \in FCtxt} r$. That is, the overall relevance to the IR condition Q of the documents that describe the facts of the *R-cube*.

Example 9. The sales shown in Table 1 constitute the set of facts F of the *R-cube*. The set of dimensions is $D = \{Products, Customers, Time, Amount\} \cup \{R, Ctxt\}$. In the previous examples we have shown the definition of some of these dimensions along with their corresponding fact-dimension relations. The IR condition used for stating the context of analysis was $Q = \text{"financial, crisis"}$. The quality of the *R-cube* is $Quality = 0.2$.

6. R-CUBES ALGEBRA

In this section we present an algebra for the *R-cubes* by extending the definition of the unary operators presented in [14] to regard the relevance and context of the facts. For each operator, we show its definition, and discuss how the relevance and context are updated in the result by giving some examples.

6.1 Selection Operator

The *selection* operator restricts the facts in the cube to the subset of facts that satisfy some given conditions.

Definition 6. Let $p : D_1 \times \dots \times D_n \times R \times Ctxt \rightarrow \{true, false\}$ be a predicate on the dimensions D . The *relevance-extended selection* operator, $R\sigma$, is defined as: $R\sigma[p](RM) = (F', D', FD', Q')$, where $F' = \{f \in F \mid \exists (e_1, \dots, e_n, r, d) \in D_1 \times \dots \times D_n \times R \times Ctxt, p(e_1, \dots, e_n, r, d) \wedge f \sim_1 e_1 \wedge \dots \wedge f \sim_n e_n \wedge f \sim_R r \wedge f \sim_{Ctx} d\}$, $D' = D$ and $FD' = \{FD'_i, i = 1 \dots n\} \cup \{FR', FCtxt'\}$. $FD'_i = \{(f', e) \in FD_i \mid f' \in F'\}$, $FCtxt' = \{(f', d'') \in FCtxt \mid f' \in F'\}$ and $FR' = \{(f', r\beta) \mid (f', r) \in FR \wedge f' \in F'\}$, $\beta = Quality/Quality' \geq 1$, $Quality' = \sum_{(f,d'') \in FCtxt'} r$ and $Q' = Q$.

The set of facts in the resulting *R-cube* is restricted to those facts characterized by the dimension values where p is true. The fact-dimension relations are restricted accordingly too. Notice that the quality of the *R-cube* will be decreased when any relevant document is discarded by the selection operation. As formally discussed in the theorem (1) of Appendix B, the relevance values are increased by a factor of β . Thus, it is ensured that the sum of the relevance values of the facts in the resulting *R-cube* remains equal to one.

| F' | Products.ProductId | Customers.Country | Time.Month | Amount | R | Ctxt |
|-------|--------------------|-------------------|------------|-----------|------|--------------------------|
| f_4 | $fo1$ | $Japan$ | 1998/10 | 300,000\$ | 0.62 | $d_1^{0.04}, d_2^{0.08}$ |
| f_5 | $fo2$ | $Korea$ | 1998/11 | 400,000\$ | 0.38 | $d_2^{0.08}, d_6^{0.01}$ |

Table 2: Result of applying $R\sigma$ on the example R -cube of Table 1, $p = (Customers.Region = Southeast Asia, R.Relevance Degree = very relevant or relevant)$. $Quality' = 1.54$.

Example 10. We can apply the relevance-extended selection operator to dice the R -cube and study the sales made to *Southeast Asian* customers. Since conditions on the relevance dimension are supported, we could also restrict the analysis to those facts considered as *relevant* or *very relevant*. Thus, $p = (Customers.Region = Southeast Asia, R.Relevance Degree = very relevant or relevant)$. Table 2 shows the resulting R -cube. The set of facts is restricted to $F' = \{f_4, f_5\}$. The resulting fact-dimension relations are: $FProducts' = \{(f_4, fo1), (f_5, fo2)\}$, $FCustomers' = \{(f_4, Japan), (f_5, Korea)\}$, $FTime' = \{(f_4, 1998/10), (f_5, 1998/11)\}$, $FAmount' = \{(f_4, 300,000\$), (f_5, 400,000\$)\}$. Notice that the stated restriction also affects the set of documents that describe the facts of the R -cube, $FCtxt' = \{(f_4, d_1^{0.04}), (f_4, d_2^{0.08}), (f_5, d_2^{0.08}), (f_5, d_6^{0.01})\}$, $FCxt' \subset FCtxt$. Since some relevant documents for the analysis context are discarded, the quality of the resulting R -cube decreases to $Quality' = 0.13 < Quality = 0.2$. Consequently, the relevance value of the facts in the result is also affected, since these values will be now normalized by $Quality'$. Thus, relevance of facts is increased in a β factor, $\beta = Quality/Quality' = 1.54$. The resulting relevance fact-dimension relation is $FR' = \{(f_1, 0.62), (f_2, 0.38)\}$.

The β factor measures the quality lost in the resulting R -cube. Good restrictions will result in low β values, since they preserve the relevant facts of the R -cube and discard the non-relevant ones. However, sometimes, we may be interested in a particular region of the cube. A high β value (a low $Quality'$) will warn the user of a meaningless result.

Example 11. When the selection operator is applied to the example R -cube of Table 1, with the predicate $p = (Customers.Region = Central America)$, the set of facts in the resulting R -cube is restricted to $F' = \{f_1\}$, and the context fact-dimension relation becomes $FCtxt' = \{d_3^{0.005}, d_7^{0.005}\}$. Consequently, the quality is reduced to $Quality' = 0.01$, resulting $\beta = 0.2/0.01 = 20$. The high β value points to a considerable quality lost, meaning that the analysis result is not significant in the selected context (e.g the financial crisis mainly affected the Southeast Asian countries).

6.2 Aggregate Formation Operator

The *aggregate formation* operator evaluates an aggregation function on the R -cube. By following [14], we assume the existence of a family of functions $g : 2^F \rightarrow D_{n+1}$ that receive a set of facts and compute an aggregation by taking the data from the requested fact-dimension relation (e.g. SUM_i takes the data from FD_i , and performs the sum).

The *Group* operator defined in [14] groups the facts characterized by the same dimension values. Given the dimension values $(e_1, \dots, e_n) \in D_1 \times \dots \times D_n$, $Group(e_1, \dots, e_n) = \{f \in F | f \rightsquigarrow_1 e_1 \wedge \dots \wedge f \rightsquigarrow_n e_n\}$.

Example 12. In the example R -cube of Table 1, we can group those sales made to Southeast Asian customers during the second half of 1998 as follows: given the dimension

values $(\top, Southeast Asia, 1998/2nd\ half, \top) \in \top_{Products} \times Region \times Half_year \times \top_{Amount}$, $Group(\top, Southeast Asia, 1998/2nd\ half, \top) = \{f_4, f_5\}$.

Definition 7. Given a new dimension D_{n+1} , an aggregation function $g : 2^F \rightarrow D_{n+1}$, and a set of grouping categories $\{C_i \in C_{D_i}, i = 1 \dots n, C_{D_i} \neq C_R, C_{Ctxt}\}$, the *relevance-extended aggregate formation* operator, $R\alpha$, is defined as $R\alpha[D_{n+1}, g, C_1, \dots, C_n](RM) = (F', D', FD', Q')$, where:

$$\begin{aligned}
F' &= \{Group(e_1, \dots, e_n) | (e_1, \dots, e_n) \in C_1 \times \dots \times C_n \\
&\quad \wedge Group(e_1, \dots, e_n) \neq \emptyset\}, \\
D' &= \{D'_i, i = 1 \dots n\} \cup \{D_{n+1}\} \cup \{R, Ctxt\}, \\
D'_i &= (C'_{D_i}, \sqsubseteq'_{D_i}), C'_{D_i} = \{C_{ij} \in C_{D_i} | C_i \sqsubseteq_{D_i} C_{ij}\}, \\
\sqsubseteq'_{D_i} &= \sqsubseteq_{D_i|C'_{D_i}}, \\
FD' &= \{FD'_i, i = 1 \dots n\} \cup \{FD_{n+1}\} \cup \{FR', FCtxt'\}, \\
FD'_i &= \{(f', e'_i) | \exists (e_1, \dots, e_n) \in C_1 \times \dots \times C_n, \\
&\quad f' = Group(e_1, \dots, e_n) \in F' \wedge e_i = e'_i\}, \\
FD_{n+1} &= \bigcup_{(e_1, \dots, e_n) \in C_1 \times \dots \times C_n} \{(Group(e_1, \dots, e_n), \\
&\quad g(Group(e_1, \dots, e_n))) | Group(e_1, \dots, e_n) \neq \emptyset\}, \\
FR' &= \{(f', r') | \exists (e_1, \dots, e_n) \in C_1 \times \dots \times C_n \\
&\quad \wedge f' = Group(e_1, \dots, e_n) \in F' \\
&\quad \wedge r' = \sum_{(f, r) \in FR, f \in Group(e_1, \dots, e_n)} r\}, \\
FCtxt' &= \{(f', d^{r'}) | \exists (e_1, \dots, e_n) \in C_1 \times \dots \times C_n \\
&\quad \wedge f' = Group(e_1, \dots, e_n) \in F' \\
&\quad \wedge d^{r'} \in \bigcup_{(f, d^r) \in FCtxt, f \in Group(e_1, \dots, e_n)} \{d^r\}\}, \\
Q' &= Q
\end{aligned}$$

Each fact in the resulting R -cube represents a group of facts of the original R -cube (those characterized by the same values in the grouping category). The aggregation function is evaluated over each group of facts and the result is stored in the new dimension D_{n+1} . The dimensions D_i, \dots, D_n are restricted to the ancestor categories of the corresponding grouping category. The $FCtxt$ fact-dimension relation now relates each new fact with the documents that were associated to any of the original facts of the corresponding group. As discussed in [15], we estimate the relevance of the facts by the frequency of their dimension values in the relevant documents. Consequently, the relevance of each group is the sum of the relevance values of the original facts in the group (see theorem (2) in Appendix B). We update the FR fact-dimension relation accordingly. Thus, the sum of the relevance values of the facts in the resulting R -cube remains

| F' | $\top_{Products}$ | $Customers'.Region$ | $Time'.Half_year$ | \top_{Amount} | $Total$ | R | $Ctxt$ |
|----------------|-------------------|------------------------|--------------------|-----------------|-------------|------|--------------------------------------|
| $\{f_1\}$ | \top | <i>Central America</i> | 1998/1st half | \top | 4,300,000\$ | 0.05 | $d_3^{0.005}, d_2^{0.005}$ |
| $\{f_2, f_3\}$ | \top | <i>Southeast Asia</i> | 1998/1st half | \top | 4,100,000\$ | 0.3 | $d_5^{0.02}, d_4^{0.04}$ |
| $\{f_4, f_5\}$ | \top | <i>Southeast Asia</i> | 1998/2nd half | \top | 700,000\$ | 0.65 | $d_1^{0.04}, d_2^{0.08}, d_6^{0.01}$ |

Table 3: Result of applying $R\alpha[Total, SUM_{Amount}, \top_{Products}, Region, Half_year, \top_{Amount}]$ on the example R -cube of Table 1.

equal to one. Likewise, the quality of the R -cube is not modified.

Example 13. In the example R -cube of Table 1, we can compute the total amount of sales per $Region$ and $Half_year$ by applying the aggregate formation operator as follows:

Let $Total = (C_{Total}, \sqsubseteq_{Total})$ be a new dimension to store the result of the sum, with the categories $C_{Total} = \{Total\}$, $\top_{Total} = Total$. Let SUM_{Amount} be the aggregation function that performs the sum of the values of the $Amount$ dimension. Since we want to evaluate the sum per $Region$ and $Half_year$, the grouping categories are $\{\top_{Products}, Region, Half_year, \top_{Amount}\}$. Table 3 shows the result of applying the aggregate formation operator $R\alpha[Total, SUM_{Amount}, \top_{Products}, Region, Half_year, \top_{Amount}]$ on the R -cube of Table 1.

In the resulting R -cube, there is a new fact for each combination (e_1, \dots, e_2) of dimension values in the given grouping categories, $(e_1, \dots, e_2) \in \top_{Products} \times Region \times Half_year \times \top_{Amount}$. In the example, the possible combinations are $(\top, Central\ America, 1998/1st\ half, \top)$, $(\top, Southeast\ Asia, 1998/1st\ half, \top)$ and $(\top, Southeast\ Asia, 1998/2nd\ half, \top)$. Each new fact represents the group of original facts characterized by the corresponding combination of grouping categories values. Thus, in the resulting R -cube, we have the facts $\{f_1\} = Group(\top, Central\ America, 1998/1st\ half, \top)$, $\{f_2, f_3\} = Group(\top, Southeast\ Asia, 1998/1st\ half, \top)$ and $\{f_4, f_5\} = Group(\top, Southeast\ Asia, 1998/2nd\ half, \top)$, obtaining $F' = \{\{f_1\}, \{f_2, f_3\}, \{f_4, f_5\}\}$.

The resulting R -cube has seven dimensions. The $Ctxt$ and R dimensions are not modified. The dimension $Products'$ and $Amount'$ have been cut to their top categories, $\top_{Products}$ and \top_{Amount} , respectively. The dimension $Customers'$ is cut, so that only the categories $Region \sqsubseteq \top_{Customers}$ are kept. The $Time'$ dimension is also cut to the categories $Half_year \sqsubseteq Year \sqsubseteq \top_{Time}$. The new dimension $Total$ stores the result of the aggregation.

The fact dimension-relations $FProducts', FCcustomers', FTime'$ and $FAmount'$, now link each new fact with the dimension values that characterize the corresponding group of original facts. For example, for the new fact $\{f_4, f_5\}$, we have that $(\{f_4, f_5\}, \top) \in FProducts'$, $(\{f_4, f_5\}, Southeast\ Asia) \in FRegion'$, $(\{f_4, f_5\}, 1998/2nd\ half) \in FTime'$ and $(\{f_4, f_5\}, \top) \in FAmount'$. The $FCtxt'$ fact dimension-relation links each new fact with the documents that were related with the original facts of the corresponding group. For example, in the original R -cube we had $\{(f_4, d_1^{0.04}), (f_4, d_2^{0.08}), (f_5, d_2^{0.08}), (f_5, d_6^{0.01})\} \subset FCtxt$, then, in the resulting R -cube we have $\{(\{f_4, f_5\}, d_1^{0.04}), (\{f_4, f_5\}, d_2^{0.08}), (\{f_4, f_5\}, d_6^{0.01})\} \subset FCtxt'$. Thus, the quality of the R -cube remains $Quality' = Quality = 0.2$. The relevance of the new facts is the sum of the relevance values of the original facts in the corresponding group. In the example, we have that $(\{f_4, f_5\}, 0.65) \in FR'$, since $\{(f_4, 0.4), (f_5, 0.25)\} \subset FR$. Finally, the new $FTotal$ fact-dimension relation links each

new fact with the result of applying the aggregation function SUM_{Amount} to the corresponding group of facts. Since $\{(f_4, 300,000\$), (f_5, 400,000\$)\} \subset FAmount$, then $(\{f_4, f_5\}, 700,000\$) \in FTtotal$.

The resulting R -cube clearly shows that the most relevant is $\{f_4, f_5\}$. That is, the financial crisis had the strongest impact in the Southeast Asian region during the second half of the year, which would explain the corresponding sales fall. We could gain insight into the context of this fact by performing a drill-through operation [19], thus retrieving the textual contents of the documents which explain the details of the crisis.

6.3 Projection Operator

The *projection operator* removes some of the cube dimensions. Next, we give the formal definition of the *relevance-extended projection operator*. It is basically the projection operator defined in [14], but restricted to avoid the removal of the *relevance* and the *context* dimensions. In this way, the R -cubes algebra here presented is closed, since the result of the three operations over R -cubes is always an R -cube.

Definition 8. Given the dimensions $D_1, \dots, D_k \in D - \{R, Ctxt\}$, the *relevance-extended projection operator*, $R\pi$, is defined as $R\pi[D_1, \dots, D_k](RM) = (F', D', FD', Q')$: $F = F'$, $D' = \{D_1, \dots, D_k\} \cup \{R, Ctxt\}$, $FD' = \{FD_1, \dots, FD_k\} \cup \{FR, FCtxt\}$, and $Q' = Q$.

Example 14. By following with the Example 13, we can apply the relevance extended-projection operator to slice the result of the aggregation by removing the $Products$ and $Amount$ dimensions. The result is equivalent to the one that would be obtained with the traditional OLAP *roll-up* operation.

Thus, by applying $R\pi[Customers, Time, Total\ Amount]$ on the R -cube of Table 3, we obtain a new R -cube with the same set of facts, the dimensions $Customers, Time, Total, R$ and $Ctxt$ as returned by the aggregation operator, along with their corresponding fact-dimension relations. Since the relevance values are not modified, the quality of the resulting R -cube remains $Quality' = Quality = 0.2$.

As discussed in [14], the *drill-down* operation is equivalent to evaluating an *aggregate formation* on lower categories. Since more detailed data is required, a reference to the original R -cube is needed.

Finally, note that an R -cube is an special *multidimensional object* [14]. Thus, an R -cube can also be operated by using the algebra proposed in the base model. In this case, the result may no longer be an R -cube, since the relevance or the context dimension may be projected away, or the facts relevance may not be updated. However, these operators can be used for performing interesting IR-like analysis to explore the documents collection. For example, the context dimension may be used as grouping category to calculate aggregations over the facts described in each document.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an architecture for the construction of contextualized warehouses. A contextualized warehouse is a new decision support system that allows users to obtain strategic information by combining all their sources of structured data and unstructured documents, and by analyzing the integrated data under different contexts. In a contextualized warehouse, the user specifies an analysis context by supplying a sequence of keywords (IR condition). The analysis is performed in an *R-cube* which is materialized by retrieving the documents and facts related to the selected context. *R-cubes*, are characterized by two special system-maintained dimensions, namely: the relevance and the context dimensions. Relevance is a numeric value that measures the importance of each fact in the established context of analysis. The context dimension relates each fact with the documents that explain its circumstances (i.e the context of the fact). We have extended an existing multidimensional data model for the *R-cubes* and studied how the relevance and context dimensions should be addressed by the unary algebra operators.

In order to validate the usefulness and performance of our approach, a prototype system is currently being built. In this prototype a structured warehouse of historical financial data is contextualized with the digital version of some popular business journals. The goal is to analyze companies stock market values and determine the causes of the increases and decreases of their stock prices. We base the implementation of the prototype on a commercial multi-dimensional database. The relevance and the context dimensions are supported by two special measures whose implicit aggregation functions are the sum and the union, respectively.

In future work, the *R-cubes* algebra should be completed with binary operators. For this purpose, data fusion mechanisms [4] can be applied to combine the relevance of the involved facts. The *R-cubes* base model supports incompleteness and imprecision [14]. We plan to exploit these properties in the future to analyze the facts described in the documents which are not available in the corporate warehouse. Another possible extension of the present work is to reflect the hierarchical structure of the XML documents in the context dimension of the *R-cubes*, so that users will be able to navigate on the structure of the XML documents in the OLAP operations. Pre-aggregation strategies for *R-cubes* also remain to be studied.

8. ACKNOWLEDGEMENTS

This work was started during a stay performed by the first author at the Database and Programming Technologies research group at the Department of Computer Science at Aalborg University.

This project has been partially funded by the Spanish National Research Project TIC2002-04586-C04-03 and the Fundación Bancaixa Castelló.

9. REFERENCES

- [1] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
- [2] K. Beyer, D. Chambérin, L. S. Colby, F. Özcan, H. Pirahesh, and Y. Xu. Extending XQuery for analytics. In *Proc. of SIGMOD*, pages 503–514, 2005.
- [3] E. F. Codd. Providing OLAP to user-analysts: An IT mandate, 1993.

- [4] W. B. Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval*, pages 1–36. Kluwer, 2000.
- [5] R. Danger, R. Berlanga, and J. Ruiz-Shulcloper. CRISOL: An Approach for Automatically Populating Semantic Web from Unstructured Text Collections. In *Proc. of DEXA*, pages 243–252, 2004.
- [6] R. Grishman. Information Extraction: Techniques and Challenges. In *Proc. of SCIE*, pages 10–27, 1997.
- [7] W. H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, 1996.
- [8] R. Kimball. *The Data Warehouse Toolkit*. John Wiley & Sons, 2002.
- [9] V. Lavrenko and W. B. Croft. Relevance-Based Language Models. In *Proc. of SIGIR*, pages 120–127, 2001.
- [10] D. M. Llidó, R. Berlanga, and M. J. Aramburu. Extracting Temporal References to Assign Document Event-Time Periods. In *Proc. of DEXA*, pages 62–71, 2001.
- [11] B. R. Moole. A Probabilistic Multidimensional Data Model and Algebra for OLAP in Decision Support Systems. In *Proc. of IEEE SoutheastCon*, 2003.
- [12] B.-K. Park, H. Han, and I.-Y. Song. XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses. In *Proc. of DaWaK*, 2005.
- [13] D. Pedersen, K. Riis, and T. B. Pedersen. XML-Extended OLAP Querying. In *Proc. of SSDBM*, pages 195–206, 2002.
- [14] T. B. Pedersen, C. S. Jensen, and C. E. Dyreson. A foundation for capturing and querying complex multidimensional data. *Inf. Syst.*, 26(5):383–423, 2001.
- [15] J. M. Pérez, R. Berlanga, and M. J. Aramburu. A Document Model Based on Relevance Modeling Techniques for Semi-structured Information. In *Proc. of DEXA*, pages 318–327, 2004.
- [16] J. M. Pérez, T. B. Pedersen, R. Berlanga, and M. J. Aramburu. IR and OLAP in XML Document Warehouses. In *Proc. of ECIR*, pages 536 – 539, 2005.
- [17] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.
- [18] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.
- [19] G. Spofford. *MDX Solutions with Microsoft SQL Server Analysis Services*. John Wiley & Sons, 2001.
- [20] W3C. Extensible Markup Language (XML) 1.0. <http://www.w3.org/TR/REC-xml>, February 2004.
- [21] L. Xyleme. A dynamic warehouse for XML data of the Web. *IEEE Data Engineering Bulletin*, 24(2):40 – 47, 2001.

APPENDIX

A. FACT RELEVANCE CALCULUS

This appendix summarizes the approach proposed in [15] to calculate the relevance of a fact to an IR condition. Given an IR condition $Q = q_1 q_2 \dots q_n$, where each q_i is a query keyword, we estimate the relevance of a fact f by calculating the probability of observing this fact in the set of documents RQ relevant to Q :

$$P(f|RQ) = \frac{\sum_{d \in RQ} P(f|d)P(Q|d)}{\sum_{d' \in RQ} P(Q|d')} \quad (1)$$

In formula (1), $P(f|d)$ is the probability of finding the fact f in a relevant document $d \in RQ$. This probability is estimated by formula (2). $P(Q|d)$ is the probability of observing the query keywords in this document. By follow-

ing [9], we assume that the query keywords q_i are independent and use formulas (3) and (4) for estimating $P(Q|d)$.

$$P(f|d) = FF(f, d)/|d|_f \quad (2)$$

$$P(Q|d) = \prod_{q_i \in Q} P(q_i|d) \quad (3)$$

$$P(q_i|d) = \lambda \frac{TF(q_i, d)}{|d|_t} + (1 - \lambda) \frac{ctf_{q_i}}{coll_size_t} \quad (4)$$

In formula (2), $FF(f, d)$ returns the frequency of the fact f dimension values in the document d . That is, the number of times that the fact f dimension values occur in the document d . $|d|_f$ is the total number of dimension values found in the document d . Finally, in formula (4), $TF(q_i, d)$ returns the frequency of the query keywords q_i in the document d (the number of occurrences of q_i in d), $|d|_t$ is the total number of words in the document d , ctf_{q_i} is the number of times that q_i occurs in all the documents of the warehouse, and $coll_size_t$ is the total number of words in all the documents of the warehouse. The λ factor is called the *smoothing parameter* [9], since it avoids probabilities equal to zero when a document does not contain all the query keywords.

Intuitively, the denominator of formula (1) measures how good the set of relevant documents RQ is, since it computes the sum of the probabilities of observing the query keywords in each document of RQ . We propose formula (5) as a measure of the quality of an R -cube:

$$Quality = \sum_{d \in RQ} P(Q|d) \quad (5)$$

B. THEOREMS AND PROOFS

This appendix presents the theorems referenced along the paper.

THEOREM 1. *Let $F = \{f\}$ be a set of facts and RQ the set of documents which describe these facts. Let $F' \subseteq F$, and $RQ' \subseteq RQ$ the documents which describe the facts in F' . The relevance $P(f|RQ')$ of the facts $f \in F'$ can be calculated as follows:*

$$P(f|RQ') = P(f|RQ)\beta, \beta = \frac{\sum_{d \in RQ} P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} \geq 1$$

PROOF. Let $f \in F'$, as discussed in Appendix A we estimate its relevance $P(f|RQ')$ by:

$$P(f|RQ') = \frac{\sum_{d \in RQ'} P(f|d)P(Q|d)}{\sum_{d' \in RQ'} P(Q|d')}$$

Since the documents $d \in RQ - RQ'$ do not describe any fact of F' , the probability of observing a fact $f \in F'$ in a node $d \in RQ - RQ'$ is $P(f|d) = 0$. Thus, we can write:

$$P(f|RQ') = \frac{\sum_{d \in RQ} P(f|d)P(Q|d)}{\sum_{d' \in RQ'} P(Q|d')}$$

The relevance $P(Q|d)$ of the documents $d \in RQ \supseteq RQ'$ do not change since the IR condition Q is maintained. In this

way, $\beta = \frac{\sum_{d \in RQ} P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} \geq 1$ (notice that $|RQ| \geq |RQ'|$). Finally, the previous formula can be expressed as:

$$\begin{aligned} P(f|RQ') &= \frac{\sum_{d \in RQ} P(f|d)P(Q|d)}{\sum_{d' \in RQ} P(Q|d)} \frac{\sum_{d' \in RQ} P(Q|d)}{\sum_{d' \in RQ'} P(Q|d)} \\ &= P(f|RQ)\beta \end{aligned}$$

□

THEOREM 2. *Let $\{C_i \in C_{D_i}, i = 1 \dots n\}$ be a set of grouping categories, and let $Group(e_1, \dots, e_n)$ be the group of facts of the cube characterized by the category values $(e_1, \dots, e_n) \in C_1 \times \dots \times C_n$. The relevance value of the group $P(Group(e_1, \dots, e_n)|RQ)$ is determined by the following formula:*

$$P(Group(e_1, \dots, e_n)|RQ) = \sum_{f_i \in Group(e_1, \dots, e_n)} P(f_i|RQ)$$

PROOF. Consider the fact f characterized by the dimension values (e_1, \dots, e_n) . By applying the formula (2), the probability $P(f|d)$ of finding the fact f in the document d can be estimated as follows:

$$\begin{aligned} P(f|d) &= \frac{FF(f, d)}{|d|_f} = \sum_{f_i \in Group(e_1, \dots, e_n)} \frac{FF(f_i, d)}{|d|_f} \\ &= \sum_{f_i \in Group(e_1, \dots, e_n)} P(f_i|d) \end{aligned}$$

That is, $P(f|d)$ can be calculated by adding the dimension values frequency of each fact of $Group(e_1, \dots, e_n)$ in the document d . Notice that $\forall f_i \in Group(e_1, \dots, e_n), f_i \rightsquigarrow_1 e_1 \wedge \dots \wedge f_i \rightsquigarrow_n e_n$.

Finally, with the previous result, the fact relevance calculus formula (1) can be expressed as:

$$\begin{aligned} P(f|RQ) &= \frac{\sum_{d \in RQ} P(f|d)P(Q|d)}{\sum_{d' \in RQ} P(Q|d')} \\ &= \sum_{d \in RQ} \frac{\left(\sum_{f_i \in Group(e_1, \dots, e_n)} P(f_i|d) \right) P(Q|d)}{\sum_{d' \in RQ} P(Q|d')} \\ &= \sum_{f_i \in Group(e_1, \dots, e_n)} \left(\frac{\sum_{d \in RQ} P(f_i|d)P(Q|d)}{\sum_{d' \in RQ} P(Q|d')} \right) \\ &= \sum_{f_i \in Group(e_1, \dots, e_n)} P(f_i|RQ) \end{aligned}$$

□