

# Contextualizing data warehouses with documents

Juan Manuel Pérez-Martínez <sup>a,\*</sup>, Rafael Berlanga-Llavori <sup>a,1</sup>,  
María José Aramburu-Cabo <sup>a,1</sup>, Torben Bach Pedersen <sup>b,2</sup>

<sup>a</sup> Jaume I University, Spain

<sup>b</sup> Aalborg University, Denmark

Available online 7 February 2007

## Abstract

Current data warehouse and OLAP technologies are applied to analyze the structured data that companies store in databases. The context that helps to understand data over time is usually described separately in text-rich documents. This paper proposes to integrate the **traditional corporate data warehouse** with a document warehouse, resulting in a contextualized warehouse. Thus, the user first selects an analysis context by supplying some keywords. Then, the analysis is performed on a novel type of OLAP cube, called an R-cube, which is materialized by retrieving and ranking the documents and corporate facts related to the selected context. © 2006 Elsevier B.V. All rights reserved.

**Keywords:** OLAP; Text-rich XML documents; Information retrieval

## 1. Introduction

Current data warehouse and OLAP 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 are highly valuable information that can help companies analyze their data. Because XML has become the standard for data exchange over the Internet [25], nowadays it is easy to find some of these documents in XML formats. Furthermore, existing XML tagging techniques [22] can be applied to give some structure to plain documents by identifying the different document sections, and exportation tools from most proprietary systems to XML-like formats are available now.

In this paper we present an architecture that integrates a corporate warehouse of structured data with a warehouse of text-rich XML documents. The resulting *contextualized warehouse* is a new type of decision support system that allows users to obtain strategic information by analyzing data under different contexts. For example, if we have a document warehouse with financial news articles, we can analyze the evolution of the sales measures of our corporate warehouse in the

\* Corresponding author. Fax: +34 964 728435.

E-mail addresses: [JuanMa.Perez@lsi.uji.es](mailto:JuanMa.Perez@lsi.uji.es) (J. Manuel Pérez-Martínez), [berlanga@lsi.uji.es](mailto:berlanga@lsi.uji.es) (R. Berlanga-Llavori), [aramburu@icc.uji.es](mailto:aramburu@icc.uji.es) (M.J. Aramburu-Cabo), [tbp@cs.aau.dk](mailto:tbp@cs.aau.dk) (T.B. Pedersen).

<sup>1</sup> Fax: +34 964 728435.

<sup>2</sup> Fax: +45 9815 9889.

context of a period of crisis described by the relevant news. Thus, it is easier to find out which products were affected by the crisis.

Here, we define a context *as a set of textual fragments that can provide analysts with strategic information important for decision-making tasks*. Since the document warehouse may contain documents about many different topics, we apply modern Information Retrieval (IR) [2] techniques to select the context of analysis from the document warehouse. 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 linked to the set of relevant documents that describe its context. In this paper we extend an existing multidimensional data model to represent these two new dimensions (relevance and context), and we study how the traditional OLAP operations are affected by them.

The relevance and context dimensions provide information about facts that can be very useful for analysis tasks. 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 to restrict the analysis to the facts described in a given subset of documents (e.g., the most relevant documents). Second, the user will be able to gain insight into the circumstances of a fact by retrieving its related documents.

The main contributions of this paper are: (1) an architecture that integrates a corporate warehouse with a document warehouse, resulting in a *contextualized warehouse*; (2) a formal definition of the multidimensional data model and the unary algebra operations to manage R-cubes; and (3) a prototypical system that shows the usefulness of the approach.

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

## 2. Related work

In [9] the importance of external contextual information to understand the results of historical analysis operations was emphasized: “External contextual infor-

mation is information outside the corporation that nevertheless plays an important role in understanding information over time.” Since contextual information is usually available as documents (e.g., on-line news, company reports, etc.), which cannot be managed by relational systems, *few approaches regarding contextual information in a data warehouse can be found in the literature. With the emergence of XML as the lingua franca of the Web, semi-structured information is now widely available, and several methods have been proposed to combine XML and data warehouses.*

The problem of gathering and querying web data is not trivial, mainly because data sources are dynamic and heterogeneous. In this context, some works are focused on the construction of repositories for XML [26] and web documents [4]. The main issues addressed by them include efficient storage, indexing, query processing, data acquisition, change control and schema integration of data extracted from heterogeneous web sources.

In [17] OLAP operations are extended to involve dimensions and/or measures coming from external XML data. Unfortunately, this approach only deals with highly-structured XML data, thus being unsuitable for text-rich XML documents. Other approaches propose to apply OLAP-like operators to aggregate information of XML structured documents. For example, [3] proposes to extend XQuery [25] with grouping constructs to evaluate OLAP-style aggregation queries on XML documents, and [15] provides mechanisms to perform text aggregations on the XML textual contents. Nevertheless, these approaches do not regard the factual data included in XML texts, and they lack the mechanisms to relate corporate data warehouses with external XML documents.

*The work presented in [21] proposes to annotate external information sources (e.g., documents, images, etc.) by means of an ontology that comprises all the values of the data warehouse’s dimensions.* In this way, each OLAP report can be associated with the external sources annotated with the same dimension values. This approach has several limitations that have been solved in our work. First, it is necessary to manually annotate all the documents, which is unfeasible for large collections. Second, this approach does not provide any mechanisms to actually integrate the corporate cubes with their contexts. Third, it does not provide a formal framework for calculating both document and fact relevance with respect to user queries.

Nowadays, any application required to manipulate large collections of documents applies *Information Retrieval technology* [2]. Recent proposals in the field of IR include Language Modeling [20] and Relevance

Modeling [11]. Language Modeling represents each document as a language model. Thus, documents are ranked according to the probability of obtaining the query keywords when taking random samples from their corresponding language models. Relevance Modeling estimates the joint probability of the query's keywords over the set of documents deemed relevant for that query. Language and Relevance Modeling outperform traditional IR models in many cases. One of the current hot topics in IR research is retrieval of XML data [5,8]. These approaches combine both keyword-based and structural retrieval conditions.

Our approach relies on Relevance Modeling mainly because of two reasons. First, Relevance Modeling provides a formal background based on the Probability Theory, which is also well-suited for OLAP operations. Second, contrary to Language Modeling, Relevance Modeling deals with sets of relevant documents instead of single documents, which seems more appropriate for representing the contexts of the facts in a data warehouse.

This paper is based on the results of previous work by the authors. In [18] we presented a model for text-rich XML documents. Over this model, several information extraction techniques have been developed to identify the facts described in the documents [7,12]. In [18] we also showed how to apply Relevance Modeling to estimate the relevance of facts extracted from documents, and in [19] we outlined its multi-dimensional implementation. In this paper we propose an approach to contextualizing corporate cubes with the facts extracted from documents, resulting in a new multi-dimensional model called R-cube.

### 3. Contextualized warehouses

A contextualized warehouse is a decision support system that allows users to combine all their sources of structured and unstructured data, and to analyze the integrated data under different contexts. Fig. 1 shows the proposed architecture for the contextualized warehouse. Its main components are a corporate warehouse, a document warehouse and the fact extractor module. The corporate warehouse is a traditional data warehouse that integrates the company's structured data sources [9,10]. The unstructured data coming from external and internal sources are stored in the document warehouse as XML documents. The fact extractor module relates the facts of the corporate warehouse with the documents that describe their contexts. In a contextualized warehouse, the user specifies an analysis context by supplying a sequence of keywords. The analysis is performed on a new type of OLAP cube, called an R-cube, which is materialized by retrieving the docu-

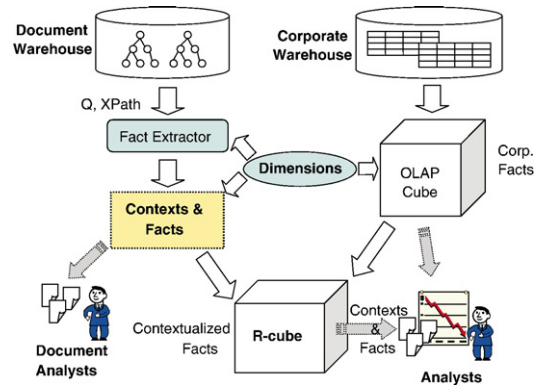


Fig. 1. Contextualized warehouse architecture.

ments and facts related to the selected context. In this section, we first present an IR model for the document warehouse, then we describe the fact extractor module, and finally we show the R-cubes construction process.

#### 3.1. An IR model for the document warehouse

The document warehouse is a repository of text-rich XML documents containing relevant information for analysis purposes. These documents are collected from the company internal and external unstructured information sources. This paper does not address the problems of data acquisition, filtering, change control and schema integration of XML data extracted from heterogeneous sources. These are studied in other works like [26]. In this section we adapt the IR model proposed in [18] to retrieve the documents that describe the analysis context. Our model, uses the traditional tree representation of XML documents and maps the elements of the original documents into nodes of the corresponding document trees.

**Definition 1.** Let  $Col = \{d\}$  be the set of all the document nodes  $d$  of all the documents in the document warehouse. With  $d' \sqsubset d$  we denote that the node  $d'$  is a descendant of  $d$ . Let  $Text(d)$  represent the sequence of words contained in all the nodes under  $d$ .

In the document warehouse, a query is a tuple  $(XPath, Q)$ , where:  $XPath$  is a path expression [25] that states a restriction on the structure of the documents; and  $Q = q_1 q_2 \dots q_n$  is an IR condition, consisting of a sequence of keywords  $q_i$ .

We define  $XPath(d)$  as a boolean function over the set of document nodes of the warehouse,  $XPath: Col \rightarrow \{True, False\}$ .  $XPath(d)$  returns *True* if the document node  $d$  is selected by the path expression  $XPath$ , and *False* otherwise.

The query ( $XPath, Q$ ) returns the set  $RQ$  of the document nodes that maximize the Relevance to the IR condition  $Q$ , which is defined as follows:

$$RQ = \{d \in Col | XPath(d) \wedge |Text(d) \cap Q| \geq m \\ \wedge \nexists d' \sqsubset d (P(Q|d') \geq P(Q|d))\}$$

Thus,  $RQ$  is the set of document nodes that are selected by the  $XPath$  expression, contain at least  $m$  query keywords and maximize the relevance with respect to their subtrees.

The relevance of the document node  $d$  to the IR condition  $Q$  is calculated by the probability  $P(Q|d)$  of observing the query keywords in the document node. By following [11], we assume that the query keywords  $q_i$  are independent, and use formulas (1) and (2) for calculating  $P(Q|d)$ .

$$P(Q|d) = \prod_{q_i \in Q} P(q_i|d) \quad (1)$$

$$P(q_i|d) = \lambda \frac{TF(q_i, d)}{|d|_t} + (1-\lambda) \frac{ctf_{q_i}}{coll\_size_t} \quad (2)$$

In formula (2),  $TF(q_i, d)$  returns the frequency of the query keywords  $q_i$  in  $Text(d)$  (the number of occurrences of  $q_i$  in  $d$  and all its child nodes),  $|d|_t$  is the total number of words in  $Text(d)$ ,  $ctf_{q_i}$  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  $\lambda$  factor is called the *smoothing parameter* [11], as it avoids probabilities equal to zero when a document does not contain all the query keywords.

The approach described above ensures that the document nodes in the result have the proper granularity level to describe the IR condition  $Q$ . For example, let ( $XPath, Q$ ) be a query in the document warehouse. First, the path expression  $XPath$  selects a subset of the document subtrees of the warehouse. Let  $d$  be a document node representing an article, and let  $d' \sqsubset d$  be a node depicting the second paragraph of this article. Both  $d$  and  $d'$  were selected by  $XPath$ . Let us consider that  $d'$  is more relevant than  $d$  for the given IR condition  $Q$ , i.e.,  $P(Q|d) < P(Q|d')$ . This setting could happen, for example, when all the query keywords only occur in the second paragraph of the article. Thus,  $d'$  and  $d$  will have the same frequencies for the query keywords. That is,  $TF(q_i, d) = TF(q_i, d') \forall q_i \in Q$ . However, the article  $d$  comprises all the words contained in  $d'$  plus all the words of the rest of paragraphs. Then,  $|d|_t > |d'|_t$ ,  $P(q_i|d) < P(q_i|d') \forall q_i \in Q$ , see formula (2), and  $P(Q|d) < P(Q|d')$ , see formula (1). Since the second paragraph is actually the document portion that better describes the informa-

tion required by  $Q$  (i.e., it obtains the maximum relevance in the subtree), the entire article  $d$  will not be included in  $RQ$ , but we will instead insert the more specific and relevant paragraph  $d'$ .

In the rest of the paper we will use the term “document” to mean “document node”, as the document nodes returned by a query are just text fragments describing the context under analysis.

### 3.2. The fact extractor module

Building a contextualized warehouse mainly means relating each fact of the corporate warehouse to its context. The fact extractor tool uses the dimensions defined in the corporate warehouse to detect the facts described in the documents. Next, we describe this process in detail by means of an example.

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. Fig. 2 shows a fragment of an example document of this warehouse. The document 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 context’s descriptions are very useful, as they contain detailed information about the facts of the corporate warehouse. For example, the document in Fig. 2 could help us to understand a sales drop.

By applying specific information extraction techniques [12,7], and considering the three analysis dimensions of the 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 for 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 completely *precise* and belong to non-base dimension categories. For example, the Southeast Asia 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 calculates the number of times that its dimension values occur in the document fragment (i.e., the fact dimension values frequency). This frequency value determines the importance of the fact in the document, and will be used to estimate the relevance of the fact.

It is worth mentioning that the fact extractor module also regards synonyms (e.g., aliment and food) and other terms semantically related to the dimension values of the corporate warehouse, in order to identify valid facts. Our current implementation of the fact extractor module is an adaptation of the work [7], which is aimed at identifying complex instances from texts to populate an ontology. In our case, the ontology is formed by the corporate dimensions, and the instances are the multidimensional facts.

Let us now consider the second sentence of the example document in Fig. 2. It depicts two facts: (Company=Chicken SPC, Time.Year=1997, Export=\$10,100,00), (Company=Chicken SPC, Time.Half\_year=1998/2nd half, Export=\$1,300,00). 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 or in external databases. However, these new facts will not be identified by our fact extraction module, since our process is guided by the corporate warehouse schema (i.e., the Company dimension is not available in the schema). Furthermore, most often 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 has been applied (i.e., average, sum, etc.) Alternatively, the system could ask the user to determine the aggregation function by showing the document contents. In this context,

different IR and information extraction-based methods for integrating documents and databases are discussed in [1]. Specifically, [1] proposes a strategy to extract from documents information related to (but not present in) the facts of the warehouse.

### 3.3. Building R-cubes

In this section we explain how the analysis cubes are materialized from the contextualized warehouse. We call them R-cubes and they include two special system-maintained dimensions, namely the *relevance* and *context* dimensions.

In order to create an R-cube the analyst must supply a query of the form ( $Q$ ,  $XPath$ ,  $MDX$ ), which states the following restrictions: ( $XPath$ ,  $Q$ ) is the query for retrieving a context from the document warehouse, and  $MDX$  are conditions over the dimensions and measures of analysis [24].

The query process takes place as follows:

1. The IR condition  $Q$  and the **path expression**  $XPath$  are evaluated in the documents' warehouse, obtaining the set of relevant documents  $RQ$ .
2. 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 only the document fragments in  $RQ$ .
3. Next, or in parallel to steps (1) and (2), the  $MDX$  conditions are evaluated on the corporate warehouse.
4. Then, each document is assigned to those facts whose dimension values can be "rolled-up" or "drilled-down" to some (possibly imprecise or incomplete) fact described by the document.
5. Finally, the relevance of each fact is calculated, resulting in an R-cube.

Continuing the running example, let us consider the analysis of the sales of food products under the context of a financial crisis reported by the business articles of the document warehouse. Thus, given  $Q$ ="financial, crisis",  $XPath$ ="/business\_newspaper/economy/article/" 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. This R-cube includes 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,  $Ctxt$ ).

```
<business_newspaper date="Dec.1,1998">
<economy>
<article>
<headline>Financial Crisis Hits Southeast Asian Market</headline>
...
<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> ...
</economy> ...
</business newspaper> ...
```

Fig. 2. Example fragment of a business journal.

As Table 1 shows, the relevance 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. We could obtain the details described in the documents by performing a *drill-through* operation on the context dimension [24]. By studying these documents we can find out that the Southeast Asian financial crisis reported by the document of Fig. 2, is a valid explanation for the sales drop. Each document  $d_i$  of the context dimension has also associated a relevance value (represented by the superscript) which measures how this document describes the analysis context.

Unlike OLAP-XML federations like those proposed in [17], 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 the system. The main advantage of this approach is that pre-aggregations can be performed over R-cubes, enabling fast analysis operations.

### 3.3.1. Fact relevance calculus

Next, we summarize the approach presented in [18] to calculate the relevance of a fact with respect to an IR condition. Intuitively, a fact will be relevant for a context if the fact is found in a document which is also relevant for the context.

Given the set of relevant documents  $RQ$  returned by the document warehouse, the relevance of a fact is estimated as the probability of observing it in  $RQ$ :

$$P(f|RQ) = \frac{\sum_{d \in RQ} P(f|d)P(Q|d)}{\sum_{d \in RQ} P(Q|d)} \quad (3)$$

In formula (3),  $P(f|d)$  is the probability of finding the fact  $f$  in a relevant document  $d \in RQ$ . This

probability is estimated by formula (4). As we defined in Section 3.1,  $P(Q|d)$  is the probability of observing the query keywords in this document (see formula (1)).

$$P(f|d) = FF(f, d)/|d|_f \quad (4)$$

In formula (4),  $FF(f, d)$  returns the frequency of the fact  $f$ 's dimension values in the document  $d$ . That is, the number of times that the dimension values of the fact  $f$  occur in the document  $d$ .  $|d|_f$  is the total number of dimension values found in the document  $d$ .

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 although all document collections are not suitable for every analysis tasks (e.g., an analysis on a financial crisis with a document collection about products manufacturing processes), the sum of the facts relevance values will be kept equal to one.

The denominator of formula (3) measures the overall relevance of the documents that satisfy the IR condition  $Q$ , that is, the sum of the probabilities of observing the query keywords in each document of  $RQ$ . Thus, we propose formula (5) as a measure of the quality of an R-cube for the selected context:

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

## 4. A multidimensional data model for R-cubes

In this section we define a formal data model for the R-cubes. We extend an existing multidimensional model [16] 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.

Table 1  
Example R-cube

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

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  $r$ .

#### 4.1. 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 [16] everything that characterizes a fact is considered to be a dimension, even those attributes modeled as measures in other approaches. Fig. 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 specific granularity level. We will write  $e \in D$ , meaning that  $e$  is a dimension value of  $D$ , if  $e \in \cup_j C_j$ .

There are two special categories present in all dimensions:  $\top_D$  and  $\perp_D \in C_D$  (the top and bottom categories). The category  $\perp_D$  has the values with the finest granularity. All 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  $\sqsubseteq_D$  on dimension values is generalized to relate dimension categories as follows: given  $C_1, C_2 \in C_D$ , then if  $C_1 \sqsubseteq_D C_2$  if  $\exists e_1 \in C_1, e_2 \in C_2, e_1 \sqsubseteq_D e_2$ . We will write  $\sqsubseteq$  and  $\leq$  instead of  $\sqsubseteq_D$  and  $\leq_D$  when it is clear that  $\sqsubseteq$  and  $\leq$  represent the partial order of the dimension  $D$ .

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

##### 4.1.1. 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:  $\perp_R=\text{Relevance} \leq \top_R$ . 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 [13], we propose to introduce an intermediate category 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.

As the relevance values are normalized to sum to one, 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  $\sqsubseteq_R^\gamma$  to map the values  $r$  of the base Relevance category to values of the Relevance Degree category depending on the value of  $r/\gamma$ . We will use  $\gamma$  as a normalization factor. Note that  $\gamma$  should measure the global relevance of a particular result. Typical measures are  $\gamma=\text{MAX}(r)$ ,  $\gamma=\text{AVG}(r)$  or  $\gamma=\text{Quality}$ .

**Definition 2.** The relevance dimension is a two-tuple  $R=(C_R, \sqsubseteq_R^\gamma)$  where:  $C_R=\{\text{Relevance, Relevance Degree, } \top_R\}$  is the set of categories;  $\text{Relevance}=[0,1]$  is the base category  $\perp_R$ ;  $\text{Relevance Degree} \in \wp([a,b])$  is a partition of the interval of Real numbers  $[a,b]$ ; and  $\sqsubseteq_R^\gamma$  is the partial order  $r \sqsubseteq_R^\gamma rd$ , if  $r \in \text{Relevance}$ ,  $rd \in \text{Relevance Degree}$  and  $r/\gamma \in rd$ .

**Example 3.** Let us consider  $\gamma=\text{MAX}(r)$  (the maximum value of relevance obtained in the R-cube), and five

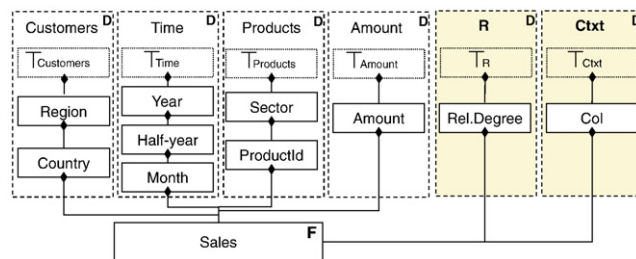


Fig. 3. Dimensions of the example case of study.

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)}, which define a partition of [0,1]. In the example of Table 1  $MAX(r)=0.4$ , 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.

#### 4.1.2. The context dimension

The context of each fact is detailed by the documents of the warehouse. We represent these documents in the context dimension.

**Definition 3.** 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  $\perp_{Ctxt}=Col=\{d\}$  is the set of the documents  $d$  of the warehouse.

**Example 4.** In our example,  $\{d_1^{0.04}, d_2^{0.08}, d_3^{0.005}, d_4^{0.04}, d_5^{0.02}, d_6^{0.01}, d_7^{0.005}\} \subset Ctxt$  are the documents of the warehouse which describe the context of the facts presented in the R-cube. The superscript denotes the relevance  $P(Q|d)$  of the document  $d$  to the context of analysis (the IR condition  $Q$ ).

The context dimension as defined in Definition 3 is flat, i.e., it has no hierarchies. It would be possible to define a hierarchy for the context dimension by considering the hierarchical structure of the XML documents. However, the IR model of the document warehouse returns the document fragments that maximize the relevance with respect to the selected context. As a consequence, the document warehouse always returns the documents at the optimal granularity level and there is no need to define a hierarchy for the context dimension.

#### 4.2. Fact-dimension relations

The fact-dimension relations link facts with dimension values. Following [16], 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 is characterized by the dimension value  $e$ , written  $f \rightsquigarrow_D e$ , 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  $f_3$ ,  $f_3 \rightsquigarrow_{Customers} Korea$  and  $f_3 \rightsquigarrow_{Customers} Southeast Asia$ .

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

#### 4.2.1. The relevance fact-dimension relation

The relevance fact-dimension relation links each fact with its relevance value.

**Definition 4.** 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  $rd \in$  Relevance Degree and  $\gamma$ , we will write  $f \rightsquigarrow_R^{rd} rd$ , meaning that the relevance degree of the fact  $f$  is  $rd$  when global relevance measure  $\gamma$  is applied, if  $\exists r \in R, (f,r) \in FR$  and  $r \sqsubseteq_R^{rd} rd$ .

**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  $\gamma=MAX(r)=0.4$ ,  $f_1 \rightsquigarrow_R^{0.4}$  very irrelevant,  $f_2 \rightsquigarrow_R^{0.4}$  irrelevant,  $f_3 \rightsquigarrow_R^{0.4}$  neutral,  $f_4 \rightsquigarrow_R^{0.4}$  very relevant and  $f_5 \rightsquigarrow_R^{0.4}$  relevant. That is,  $f_5$  is relevant, but  $f_2$  may be irrelevant for the selected context.

#### 4.2.2. The context fact-dimension relation.

The context fact-dimension relation links each fact with the documents that describe its context.

**Definition 5.** We define the context fact-dimension relation as the set  $FCtxt=\{(f,d)\}$  where  $f \in F$  is a fact described by the document  $d \in Ctxt$ , also written  $f \rightsquigarrow_{Ctxt} d$ . We denote by  $RQ$  the set of documents relevant for the analysis that describe the facts in  $F$ ,  $RQ=\bigcup_{(f,d) \in FCtxt} \{d\}$ .

**Example 8.** In the example,  $FCtxt=\{(f_1, d_3^{0.005}), (f_1, d_7^{0.005}), (f_2, d_5^{0.02}), (f_3, d_4^{0.04}), (f_4, d_1^{0.04}), (f_1, d_2^{0.08}), (f_5, d_2^{0.08}), (f_5, d_6^{0.01})\}$ . Thus, the set of documents relevant for the analysis is  $RQ=\{d_1^{0.04}, d_2^{0.08}, d_3^{0.005}, d_4^{0.04}, d_5^{0.02}, d_6^{0.01}, d_7^{0.005}\}$ . The documents  $d_1^{0.04}, d_2^{0.08}$  depict the context of the fact  $f_4$ , then  $f_4 \rightsquigarrow_{Ctxt} d_1^{0.04}$  and  $f_4 \rightsquigarrow_{sub Ctxt} d_2^{0.08}$ .



### 4.3. R-cubes: relevance-extended multidimensional objects

We extend the definition of multi-dimensional object [16] to include the relevance and context dimensions discussed before.

**Definition 6.** 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 defined above; and  $Q$  is an IR condition. In the model, we represent the relevance of each fact with respect to the context established by the IR condition  $Q$ .

We measure the analysis quality of an R-cube for the selected context by  $Quality = \sum_{d \in RQ} P(Q|d)$ . 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  $D = \{\text{Products, Customers, Time, Amount}\} \cup \{R, Ctxt\}$ . 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  $Quality = 0.2$ .

## 5. The R-cubes algebra

In this section we present an algebra for the R-cubes by extending the definition of the unary operators presented in [16] 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.

Along the definitions we will assume an R-cube  $RM=(F,D,FD,Q)$ , where  $D=\{D_i, i=1, \dots, n\} \cup \{R, Ctxt\}$ ,  $FD=\{FD_i, i=1, \dots, n\} \cup \{FR, FCtxt\}$  and whose quality is  $Quality$ . The set of documents relevant for the analysis query  $Q$  in the R-cube is denoted by  $RQ$ .

### 5.1. Selection operator

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

**Definition 7.** Let  $p: D_1 \times \dots \times D_n \times R \times Ctxt \rightarrow \{\text{true, false}\}$  be a predicate on the dimensions in  $D$ . The

relevance-extended selection operator,  $\sigma_R$ , is defined as  $\sigma_R[p](RM)=(F', D', FD', Q')$ , where:

$$\begin{aligned}
 F' &= \{f \in F \mid \exists (e_1, \dots, e_n, r, d) \in D_1 \times \dots \times D_n \\
 &\quad \times R \times Ctxt (p(e_1, \dots, e_n, r, d) \wedge f \rightsquigarrow_1 e_1 \\
 &\quad \wedge \dots \wedge f \rightsquigarrow_n e_n \wedge f \rightsquigarrow_R r \wedge f \rightsquigarrow Ctxt d)\}, \\
 D' &= D, \\
 FD' &= \{FD'_i, i=1 \dots n\} \cup \{FR', FCtxt'\} \\
 FD'_i &= \{(f', e) \in FD_i \mid f' \in F'\}, \\
 FCtxt' &= \{(f', d) \in FCtxt \mid f' \in F'\}, \\
 RQ' &= \{d \mid \exists (f', d) \in FCtxt'\} \\
 FR' &= \{(f', r') \mid \exists (f', r) \in FR \wedge f' \in F' \wedge r' \\
 &\quad = \beta r + \delta(f')\} \\
 \beta &= \frac{Quality}{Quality'} \geq 1, Quality' = \sum_{d \in RQ'} P(Q|d), \\
 \delta(f') &= \sum_{\{d \in RQ' \mid \exists (f, d) \in FCtxt \wedge FCtxt'\}} \left( \frac{P(f'|d)}{Quality'} \right. \\
 &\quad \left. - \frac{P(f'|d)}{Quality'} \right) P(Q|d) \geq 0, \\
 P(f'|d)' &= \frac{FF(f', d)}{\sum_{(f, d) \in FCtxt'} FF(f, d)}, \\
 P(f'|d) &= \frac{FF(f', d)}{\sum_{(f, d) \in FCtxt} FF(f, d)}, \\
 Q' &= Q
 \end{aligned}$$

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. In particular, the documents that do not describe selected facts are removed from the  $FCtxt$  fact-dimension relation and from  $RQ$ . In this way, the quality of the R-cube will decrease if a relevant document is discarded. As formally discussed in Theorem 1 of Appendix A, the relevance values of the facts after the selection are increased by a factor of  $\beta$ . The  $\beta$  factor represents the relative increment of importance of the selected documents when other documents of the warehouse are discarded. In addition, if a fact  $f'$  is described in documents which also describe non-selected facts, its relevance is also incremented by  $\delta(f')$ . This increment represents the increase of importance of the selected fact  $f'$  in the documents, when the non-selected facts are no longer taken into account. Thus, it is ensured that the sum of the relevance values of the facts in the resulting R-cube remains equal to one.

**Example 10.** We can apply the relevance-extended selection operator to dice the R-cube to 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,  $p = (\text{Customers.Region} = \text{South-R.Relevance Degree} = \text{very relevant or relevant})$ . 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:  $F = \{(f_4, fo1), (f_5, fo2)\}$ ,  $F = \{(f_4, Japan), (f_5, Korea)\}$ ,  $F = \{(f_4, 1998/10), (f_5, 1998/11)\}$ ,  $FD_{\text{Amount}}' = \{(f_4, 300,000\$), (f_5, 400,000\$)\}$ . The stated restriction also affects the set of documents that describe the facts of the R-cube,  $FCtxt' = \{(f_4, d_1^{0.04}), (f_4, d_2^{0.08}), (f_5, d_2^{0.08}), (f_5, d_6^{0.01})\}$ , and then  $RQ' = \{d_1^{0.04}, d_2^{0.08}, d_6^{0.01}\} \subset RQ$ . Since some documents relevant for the analysis context are discarded, the quality of the resulting R-cube decreases to  $Quality' = 0.13$  ( $< Quality = 0.2$ ). Notice that all the facts related to  $d_1, d_2$  and  $d_6$  in  $FCtxt$  were selected by the operation. That is, in this case, all the documents in the resulting R-cube only describe selected facts. Consequently, the relevance of the facts is increased in a  $\beta$  factor,  $\beta = Quality / Quality' = 1.54$ . The resulting relevance fact-dimension relation is  $FR' = \{(f_4, 0.615), (f_5, 0.385)\}$ .

**Example 11.** Let us now consider the result of the previous example. If we select the sales made during the month of November 1998 (i.e.,  $p = (\text{Time.Month} = 1998/11)$ ) from the R-cube shown in Table 2, the new set of facts is  $F' = \{f_5\}$ , and the context fact-dimension relation becomes  $FCtxt' = \{(f_5, d_2^{0.08}), (f_5, d_6^{0.01})\}$ , resulting  $RQ' = \{d_2^{0.08}, d_6^{0.01}\}$  and  $Quality' = 0.09$ , then  $\beta = 0.13 / 0.09 = 1.444$ . Document  $d_6$  only describes  $f_5$ , the selected fact. However, document  $d_2$  describes both the selected fact,  $f_5$ , and the discarded one,  $f_4$ , since in the input R-cube we had that  $\{(f_4, d_2^{0.08}), (f_5, d_2^{0.08})\} \subset FCtxt$ . The relevance of  $f_5$  in the input R-cube was 0.385,  $(f_5, 0.385) \in FR$ . Then, in the resulting R-cube, the relevance of  $f_5$  will be recalculated as  $\beta 0.385 + \delta(f_5)$ . Let  $FF(f_4, d_2) = FF(f_5, d_2) = 3$ , the dimension values of the fact  $f_4$  appear three times in document  $d_2$ , likewise, the frequency of the dimensions values of the fact  $f_5$  in  $d_2$  is three. Thus, we have that  $P(f_5|d_2) = 3/(3+3) = 0.5$  and  $P(f_5|d_2)' = 3/3 = 1$ . The relevance of document  $d_2$  to the IR condition is  $P(Q|d_2) = 0.08$ . Then,  $\delta(f_5) = (P(f_5|d_2)' - P(f_5|d_2))P(Q|d_2)/Quality' = 0.444$ . In this way, we

finally have that  $\beta 0.385 + \delta(f_5) = 1$ , and the resulting relevance fact-dimension relation is  $FR' = \{(f_5, 1)\}$ .

The  $\beta$  factor measures the quality lost in the resulting R-cube. Good restrictions will result in low  $\beta$  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  $\beta$  value (a low  $Quality'$ ) will warn the user of a meaningless result.

**Example 12.** When the selection operator is applied to the example R-cube of Table 1, with the predicate  $p = (\text{Customers.Region} = \text{Central America})$ , the set of facts in the resulting R-cube is restricted to  $F' = \{f_1\}$ , the context fact-dimension relation becomes  $FCtxt' = \{(f_1, d_3^{0.005}), (f_1, d_7^{0.005})\}$  and  $RQ' = \{d_3^{0.005}, d_7^{0.005}\}$ . Consequently, the quality is reduced to  $Quality' = 0.01$ , resulting  $\beta = 0.2 / 0.01 = 20$ . The high  $\beta$  value points to a considerable lost quality, meaning that the analysis result is not significant in the selected context (as the financial crisis mainly affected the Southeast Asian countries).

## 5.2. Aggregate formation operator

The aggregate formation operator evaluates an aggregation function on the R-cube. Following [16], 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 [16] groups the facts characterized by the same dimension values. Given the 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 13.** 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  $(T, \text{Southeast Asia}, 1998/2\text{nd half}, T) \in T_{\text{Products}} \times \text{Region} \times \text{Half\_year} \times T_{\text{Amount}}$ ,  $Group(T, \text{Southeast Asia}, 1998/2\text{nd half}, T) = \{f_4, f_5\}$ .

Table 2

Result of applying  $\sigma_R$  on the example R-cube of Table 1,  $p = (\text{Customers.Region} = \text{Southeast Asia}, R.\text{Relevance Degree} = \text{very relevant or relevant})$

$F'$	Products.ProductId	Customers.Country	Time.Month	Amount	$R$	$Ctxt$
$f_4$	fo1	Japan	1998/10	300, 000\$	0.615	$d_1^{0.04}, d_2^{0.08}$
$f_5$	fo2	Korea	1998/11	400, 000\$	0.385	$d_2^{0.08}, d_6^{0.01}$

$Quality' = 1.54$ .

**Definition 8.** Given a new dimension  $D_{n+1}$ , an aggregation function  $g: 2^F \rightarrow D_{n+1}$ , and a set of grouping categories  $\{C_i \in C_D, i=1..n, C_{D_i} \neq C_R, C_{Ctxt}\}$ , the relevance-extended aggregate formation operator,  $\alpha_R$ , is defined as  $\alpha_R[D_{n+1}, g, C_1, \dots, C_n](RM) = (F', D', FD', Q')$ , where:

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

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  $D_1, \dots, D_n$  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 with any of the original facts of the corresponding group. Notice that the set of documents relevant to the analysis query  $RQ$  does not change. Likewise, the quality of the R-cube is not modified. As discussed in Section 3, we estimate the relevance of the facts by the frequency of their dimension values in the relevant documents. Conse-

quently, the relevance of each group is the sum of the relevance values of the original facts in the group (see Theorem 2 in Appendix A). We update the  $FR$  fact-dimension relation accordingly. Thus, the sum of the relevance values of the facts in the resulting R-cube remains equal to one.

**Example 14.** 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  $C_{Total} = \{Total\}$ ,  $\sqsubseteq_{Total} = Total \leq Total$ . 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  $\{T_{Products}, Region, Half\_year, T_{Amount}\}$ . Table 3 shows the result of applying the aggregate formation operator  $\alpha_R [Total, SUM_{Amount}, T_{Products}, Region, Half\_year, T_{Amount}]$  to 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,  $(e_1, \dots, e_2) \in T_{Products} \times Region \times Half\_year \times T_{Amount}$ . In the example, the possible combinations are  $(T, Central\ America, 1998/1st\ half, T)$ ,  $(T, Southeast\ Asia, 1998/1st\ half, T)$  and  $(T, Southeast\ Asia, 1998/2nd\ half, T)$ . Each new fact represents the group of original facts characterized by the corresponding combination of grouping category values. Thus, in the resulting R-cube, we have the facts  $\{f_1\} = Group(T, Central\ America, 1998/1st\ half, T)$ ,  $\{f_2, f_3\} = Group(T, Southeast\ Asia, 1998/1st\ half, T)$  and  $\{f_4, f_5\} = Group(T, Southeast\ Asia, 1998/2nd\ half, T)$ , 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 restricted to their top categories,  $T_{Products}$  and  $T_{Amount}$ , respectively. The dimension  $Customers'$  is reduced, so that only the categories  $Region \leq T_{Customers}$  are kept. The  $Time'$  dimension is also reduced to the categories  $Half\_year \leq Year \leq T_{Time}$ . The new dimension  $Total$  stores the result of the aggregation.

Table 3  
Result of applying  $\alpha_R [Total, SUM_{Amount}, T_{Products}, Region, Half\_year, T_{Amount}]$  on the example R-cube of Table 1

$F'$	$T_{Products}$	$Customers'.Region$	$Time'.Half\_year$	$T_{Amount}$	Total	$R$	$Ctxt$
$\{f_1\}$	T	Central America	1998/1st half	T	4, 300, 000\$	0.05	$d_3^{0.005}, d_7^{0.005}$
$\{f_2, f_3\}$	T	Southeast Asia	1998/1st half	T	4, 100, 000\$	0.3	$d_5^{0.02}, d_4^{0.04}$
$\{f_4, f_5\}$	T	Southeast Asia	1998/2nd half	T	700, 000\$	0.65	$d_1^{0.04}, d_2^{0.08}, d_6^{0.01}$

The fact dimension-relations  $FD_{Products'}$ ,  $FD_{Customers'}$ ,  $FD_{Time'}$  and  $FD_{Amount'}$ , 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  $(\{f_4, f_5\}, \top) \in FD_{Products'}$ ,  $(\{f_4, f_5\}, \text{Southeast Asia}) \in FD_{Customers'}$ ,  $(\{f_4, f_5\}, 1998/2\text{nd half}) \in FD_{Time'}$  and  $(\{f_4, f_5\}, \top) \in FD_{Amount'}$ . 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 aggregate formation operation never modifies the set of documents relevant for the analysis, i.e.,  $RQ' = RQ$ . Then, the quality of the R-cube remains, i.e.,  $Quality' = Quality = 0.2$ . 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  $FD_{Total}$  fact-dimension relation links each new fact with the result of applying the aggregation function  $SUM_{Amount}$  to the corresponding group of facts. Since  $\{(f_4, 300,000\$), (f_5, 400, 000 \$)\} \subset FD_{Amount}$ , then  $(\{f_4, f_5\}, 700, 000\$) \in FD_{Total}$ .

The resulting R-cube clearly shows that the most relevant fact 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 [24], thus retrieving the textual contents of the documents that explain the details of the crisis.

### 5.3. 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 [16], but restricted to avoid the removal of the relevance and the context dimensions. In this way, we can conclude that since the result of the three operations over R-cubes is always an R-cube, the R-cubes algebra presented here is closed.

**Definition 9.** Given the dimensions  $D_1, \dots, D_k \in D \setminus \{R, Ctxt\}$ , the relevance-extended projection operator,  $\pi_R$ , is defined as  $\pi_R[D_1, \dots, D_k](RM) = (F', D', FD', Q')$ :  $F = F'$ ,  $D' = \{D_1, \dots, D_k\} \cup \{R, Ctxt\}$ ,  $FD' = \{FD_1, \dots, FD_k\} \cup \{FR, FCtxt\}$ , and  $Q' = Q$ .  $RQ' = RQ$  and  $Quality' = Quality$ .

**Example 15.** By following with the Example 14, we can apply the relevance extended-projection operator to remove the Products and Amount dimensions. The result is equivalent to the one that would be obtained with the traditional roll-up operation.

Thus, by applying  $\pi_R[Customers, Time, Total]$  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  $FCtxt$  fact-dimension relation is not modified,  $RQ' = RQ$  and the quality of the resulting R-cube remains, i.e.,  $Quality' = Quality = 0.2$ .

The drill-down operation is equivalent to evaluating an aggregate formation on lower categories [16]. Since more detailed data is required, a reference to the original R-cube is needed.

Finally, note that an R-cube is an special multi-dimensional object [16]. Thus, an R-cube can also be queried by using the algebra proposed in the base model. In this case, the result may no longer be an R-cube, as the relevance or the context dimension may be projected away, or the fact relevance may not be updated. However, these operators could be applied to perform interesting analysis. For example, the context dimension may be used as a grouping category to calculate aggregations over the facts described in each document.

## 6. The prototype

In order to validate the usefulness of our approach, we have developed a prototype. The resulting contextualized warehouse allows users to analyze stock market indexes with the advantage of having each measure value associated to a news extract that explains it. This section gives an overview of the main aspects involved in the design of the system. First, we describe the document and corporate warehouses of the prototype. Afterwards, we show the usefulness of the prototype by means of an example usage case, and explain the analysis process by means of a sequence of screen-shots. Finally, we summarise some implementation issues.

The document warehouse consists of a digital collection of some well-known international business newspapers. We inserted in the prototype a total of 132 articles from the issues published during 1990. Among other things, these articles report the trends of markets during that period. It is usual to find news explaining how stock markets are affected by some financial circumstances, e.g.: “The reaction of German



market to the rise of interest rates is expected to be ...”.

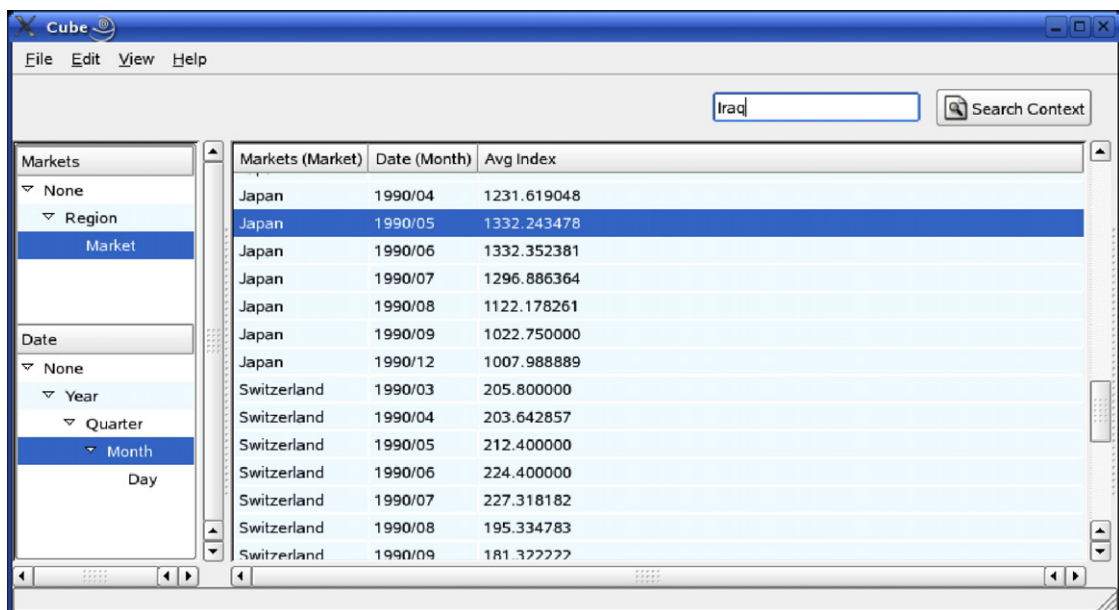
The corporate warehouse keeps a historical record of market indexes as measured by Morgan Stanley Capital International Perspective [14]. In our experiments we have only considered the indexes of the year 1990, resulting, at the lowest dimension categories, in 1396 facts. As Fig. 4 shows, the corporate cube has two dimensions. The Market dimension is organized into two categories: Market (U.S., Japan, etc.) and Region (North America, Asia, etc.). The Date dimension is organized into Day, Month, Quarter and Year. The fact (Japan, 1990/05, 1332.24) depicts that the average index in the Japanese market during May 1990 was 1332.24.

Like in traditional data warehouses, an OLAP interface allows analysts to query the corporate warehouse. Among other things, it is possible to study the average index of the different markets, pivot to order by date, roll-up to calculate the average per region, or dice the cube to select the index values of the second quarter of 1990 in Germany (see Fig. 4).

Let us suppose that there are recent news about a conflict happening in the Middle East. During the last decades conflicts have been frequent in this area, so the analyst decides to use the prototype to study the reaction of the stock markets to the Iraq war of 1990. After entering the keyword “Iraq” in the toolbar and clicking on “Search Context”, the system presents to the user a list of documents about Iraq ranked by relevance (see the left side of

Fig. 5). By selecting a document, its contents appear in the right part of the window, and the paragraph that contains the keyword is highlighted. Then, the analyst can refine the query by adding or removing keywords, and by specifying a minimum relevance threshold. It is also possible to provide some user feedback by clicking on the check boxes associated to the documents that better describe the Iraq conflict of 1990. Once the set of documents that describe the context under analysis has been obtained, the “Contextualize” button is used to continue the analysis in the OLAP window shown in Fig. 6.

Now this window presents an R-cube that includes the relevance and the context values assigned to each fact of the original cube. Dark colours depict very relevant facts, whereas light colours mark the irrelevant ones. By rolling up to the Region and Quarter levels and ordering the facts by relevance, the analyst discovers that the most relevant facts involve the Asian markets and the third quarter of 1990. Then, the analyst decides to execute a drill-down operation to study the average index per month in the Asian countries. The most relevant facts correspond to Japan and the months of August and September. As can be seen in Fig. 6, the Japanese market index had a sharp fall during these months, a fall of 100 points, whereas the average falls in the rest of markets were of about 10 points. By selecting the fact that represents the average index of the Japanese market in August, the system presents the documents that describe the context of this fact (see the right side of



Markets (Market)	Date (Month)	Avg Index
Japan	1990/04	1231.619048
Japan	1990/05	1332.243478
Japan	1990/06	1332.352381
Japan	1990/07	1296.886364
Japan	1990/08	1122.178261
Japan	1990/09	1022.750000
Japan	1990/12	1007.988889
Switzerland	1990/03	205.800000
Switzerland	1990/04	203.642857
Switzerland	1990/05	212.400000
Switzerland	1990/06	224.400000
Switzerland	1990/07	227.318182
Switzerland	1990/08	195.334783
Switzerland	1990/09	181.322222

Fig. 4. OLAP window.

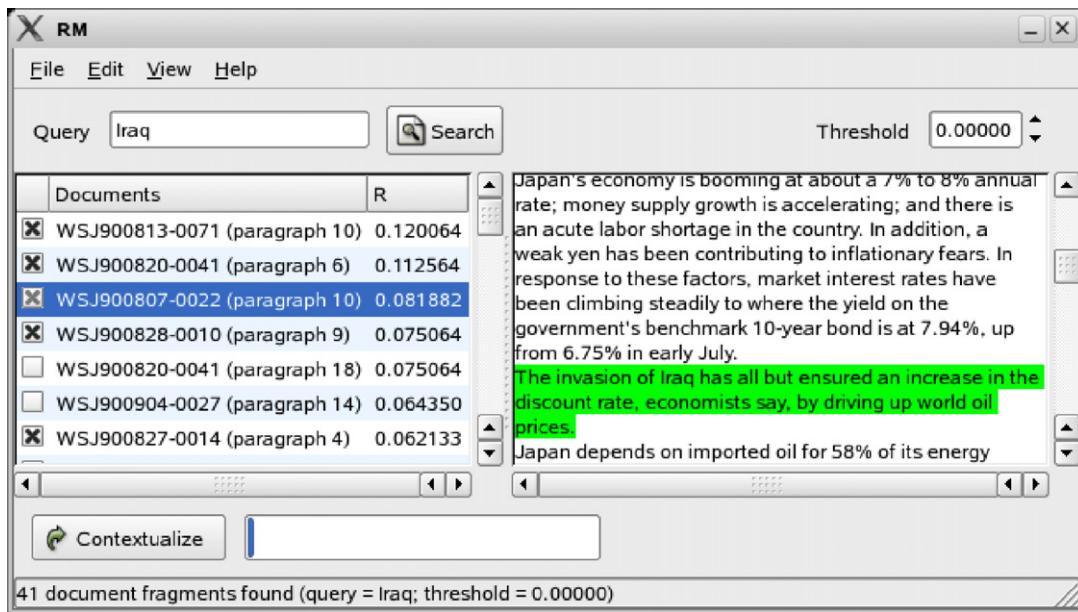


Fig. 5. IR window.

the window shown in Fig. 6). In the highlighted paragraph of the first document, the analyst discovers that “... plant engineering companies fell as their projects in Iraq and Kuwait were frozen because of the economic sanction of Japan against Iraq”. Thus, the analyst concludes that it could be a good idea to watch

Japanese investments now that there is a new conflict in the Middle East which could be as important for the financial markets as the Iraq war of 1990.

The prototype has been implemented as a set of Python modules. In order to evaluate keyword-based searches over the XML collection, the document

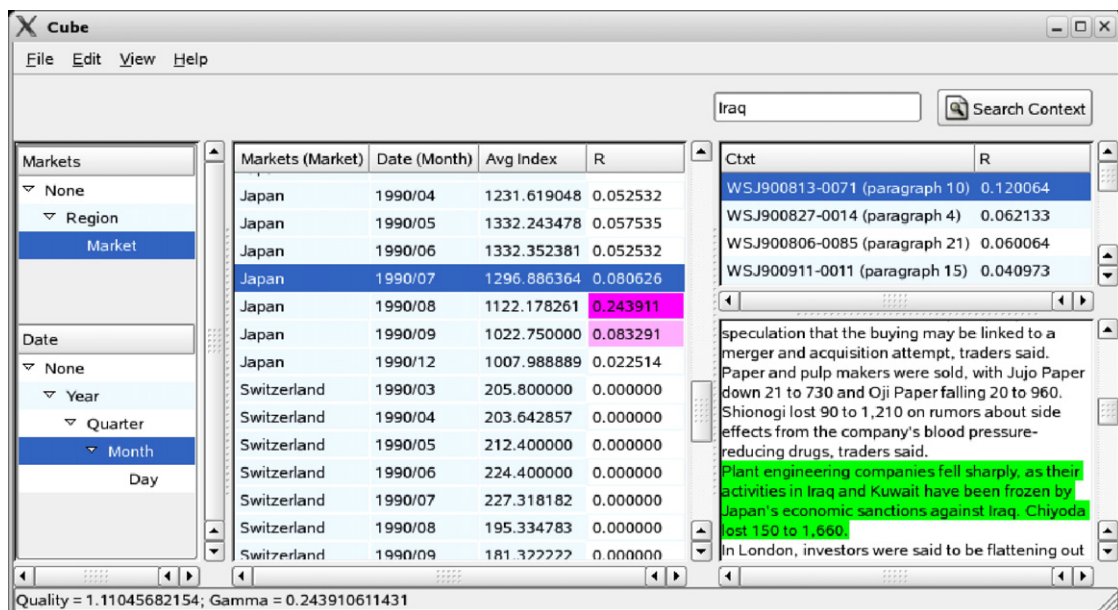


Fig. 6. OLAP window showing an R-cube.

warehouse keeps a inverted file index [2] and implements the Relevance Modelling logic of the IR model presented in Section 3.1. Stemming and proper noun recognition tasks are executed by the Tree Tagger tool [23]. The corporate cubes and OLAP operations have been supported by implementing the data model and algebra operators of the base multidimensional model [16]. The Fact Extractor module provides the methods to build the R-cube by looking for date, stock market, and region references in the paragraphs of the documents. Finally, analysis capabilities over R-cubes have been provided by implementing the data model and algebra operators discussed in this paper.

## 7. Conclusions and future work

A contextualized warehouse is a new decision support system that allows users to combine all their sources of structured data and unstructured documents, and to obtain strategic information by analyzing the integrated data under different contexts. In a contextualized warehouse, the user specifies an analysis context by supplying some keywords. Then, the analysis is performed on an R-cube which is materialized by retrieving the documents and facts related to the selected context.

R-cubes are characterized by two special dimensions, namely the relevance and context dimensions. The relevance is a numeric value that measures the importance of each fact in the context of analysis. The context dimension relates each fact with the documents that explain its circumstances. In order to formalize the definition of R-cubes, we have extended a multidimensional data model [16] and studied how the relevance and context dimensions should be addressed by the unary algebra operators. This algebra remains to be completed with binary operators. For this purpose, data fusion mechanisms [6] can be applied to combine the relevance of the involved facts.

In this paper, the usefulness of contextualized data warehouses has been shown by means of a prototype. Testing the performance of the system with larger data sets, and studying query evaluation techniques for R-cubes, like pre-aggregation strategies, will be future work.

In this work we have shown how the dimension values found in documents can be applied in the process of relating them with the corporate facts that have the same dimension values. Trying to analyze the facts extracted from the documents without considering the corresponding corporate facts is an even more challenging task. In this case, the analysis may involve facts

that are incomplete (not all the dimensions may be quoted in the documents contents) and/or imprecise (if the dimension values found belong to non-base granularity levels). The R-cubes base model supports incompleteness and imprecision [16]. For the future, we plan to exploit these features to analyze the facts described in the documents that are not available in the corporate warehouse.

## Acknowledgements

This project has been partially supported by the Danish Research Council for Technology and Production under grant no. 26-02-0277, the Spanish National Research Project TIN2005-09098-C05-04, and the Fundación Bancaixa Castelló.

## Appendix A. Theorems and proofs

**Theorem 1.** Let  $RM = (F, D, FD, Q)$  be an R-cube and  $RM' = (F', D', FD', Q')$  the R-cube obtained after applying the selection operation  $\sigma_R[p]$  over  $RM$ ,  $\sigma_R[p](RM) = RM'$ . The relevance of the facts  $f' \in F'$  can be calculated as  $P(f'|RQ') = \beta P(f'|RQ) + \delta(f')$ , where:

$$\begin{aligned} \beta &= \frac{Quality}{Quality'} \geq 1, \\ \delta(f') &= \sum_{\{d \in RQ' | \exists (f, d) \in FCtxt \setminus FCtxt'\}} \left( \frac{P(f'|d)'}{Quality'} \right. \\ &\quad \left. - \frac{P(f'|d)}{Quality'} \right) P(Q|d) \geq 0 \\ P(f'|d)' &= \frac{FF(f', d)}{\sum_{(f, d) \in FCtxt'} FF(f, d)}, \\ P(f'|d) &= \frac{FF(f', d)}{\sum_{(f, d) \in FCtxt} FF(f, d)} \end{aligned}$$

**Proof.** Let  $f' \in F'$ , as discussed in Section 3.3, we estimate its relevance  $P(f'|RQ')$  by:

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

Notice that the probability  $P(f'|d)'$  of observing the fact  $f'$  in a document  $d$  when considering the restricted set of facts  $F'$ , is different from the probability  $P(f'|d)$

of observing the fact  $f'$  in  $d$  when considering the super-set  $F$ .

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

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

Let  $RQ_1$  be the subset of documents that only describe facts in  $F'$ ,  $RQ_1 = \{d \in RQ | \nexists (f, d) \in FCtxt \setminus FCtxt'\}$ ; and  $RQ_2$  the subset of document that at least describe a fact that was in  $F$  but not in  $F'$ ,  $RQ_2 = \{d \in RQ | \exists (f, d) \in FCtxt \setminus FCtxt'\}$ . The subsets  $RQ_1$  and  $RQ_2$  as defined above constitute a partition of  $RQ$ , i.e.  $RQ_1 \cap RQ_2 = \emptyset$  and  $RQ_1 \cup RQ_2 = RQ$ , then:

$$P(f'|RQ') = \frac{\sum_{d \in RQ_1} P(f'|d)' P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} + \frac{\sum_{d \in RQ_2} P(f'|d)' P(Q|d)}{\sum_{d \in RQ'} P(Q|d)}$$

Since the documents in  $RQ_1$  only describe facts in  $F'$ , we have that  $\forall d \in RQ_1$ ,  $P(f'|d)' = \frac{FF(f', d)}{\sum_{(f, d) \in Ctxt} FF(f, d)} = \frac{FF(f', d)}{FF(f, d)}$ , and consequently:

$$P(f'|RQ') = \frac{\sum_{d \in RQ_1} P(f'|d) P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} + \frac{\sum_{d \in RQ_2} P(f'|d)' P(Q|d)}{\sum_{d \in RQ'} P(Q|d)}$$

The previous formula can be rewritten as follows:

$$P(f'|RQ') = \left( \frac{\sum_{d \in RQ_1} P(f'|d) P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} + \frac{\sum_{d \in RQ_2} P(f'|d) P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} + \frac{\sum_{d \in RQ_2} P(f'|d)' P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} - \frac{\sum_{d \in RQ_2} P(f'|d) P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} \right) \frac{\sum_{d \in RQ} P(Q|d)}{\sum_{d \in RQ} P(Q|d)}$$

Since  $RQ_1 \cup RQ_2 = RQ$ , we have that:

$$P(f'|RQ') = \frac{\sum_{d \in RQ} P(f'|d) P(Q|d)}{\sum_{d \in RQ} P(Q|d)} \frac{\sum_{d \in RQ} P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} + \frac{\sum_{d \in RQ_2} (P(f'|d)' - P(f'|d)) P(Q|d)}{\sum_{d \in RQ'} P(Q|d)}$$

The relevance  $P(Q|d)$  of the documents  $d \in RQ \supseteq RQ'$  do not change because the IR condition  $Q$  is maintained. In this way,  $\beta = \frac{Quality}{Quality'} = \frac{\sum_{d \in RQ} P(Q|d)}{\sum_{d \in RQ'} P(Q|d)} \geq 1$  (notice that  $|RQ| \geq |RQ'|$ ). On the other hand,  $P(f'|d)' \geq P(f'|d)$  because  $|\{(f, d) \in FCtxt'\}| \leq |\{(f, d) \in FCtxt'\}|$  and  $\sum_{(f, d) \in FCtxt'} FF(f, d) \leq \sum_{(f, d) \in FCtxt} FF(f, d)$ . Finally, the previous formula can be expressed as:

$$P(f'|RQ') = \beta P(f'|RQ) + \delta(f') \\ \delta(f') = \sum_{\{d \in RQ' | \exists (f, d) \in FCtxt \setminus FCtxt'\}} \left( \frac{P(f'|d)'}{Quality} - \frac{P(f'|d)}{Quality'} \right) P(Q|d) \geq 0$$

□

**Theorem 2.** Let  $\{C_i \in C_D, i = 1..n\}$  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 (4), the probability  $P(f|d)$  of finding the fact  $f$  in the document  $d$  can be estimated as follows:

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

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  $\forall f_i \in Group(e_1, \dots, e_n), f_i \rightsquigarrow e_1 \wedge \dots \wedge f_i \rightsquigarrow e_n$ .



Thus, with the previous result, the fact relevance calculus formula (3) can be expressed as:

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

## References

- [1] A. Badia, Text warehousing: present and future, in: J. Darmont, O. Boussaid (Eds.), *Processing and Managing Complex Data for Decision Support*, Idea Group, 2006, pp. 96–121.
- [2] R.A. Baeza-Yates, B.A. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press/Addison-Wesley, 1999.
- [3] K. Beyer, D. Chabérin, L.S. Colby, F. Özcan, H. Pirahesh, Y. Xu, Extending XQuery for analytics, *Proc. of SIGMOD*, ACM Press, New York, 2005, pp. 503–514.
- [4] S. Bhowmick, S.K. Madria, W.-K. Ng, E.-P. Lim, Web warehousing: design and issues, *Proc. of DWDW*, Springer-Verlag, London, 1998, pp. 93–104.
- [5] T.T. Chinenyanga, N. Kushmerick, An expressive and efficient language for XML information retrieval, in: G. Mecca, J. Siméon (Eds.), *Proc. of WebDB*, ACM Press, New York, 2001, pp. 1–6.
- [6] W.B. Croft, Combining approaches to information retrieval, *Advances in Information Retrieval*, Kluwer Academic Publishers, Boston, 2000, pp. 1–36.
- [7] R. Danger, I. Sanz, R. Berlanga, J. Ruiz-Shulcloper, A proposal for the automatic generation of instances from unstructured text, in: J.F. Martínez, J.A. Carrasco-Ochoa (Eds.), *Proc. of CIARP*, Springer-Verlag, 2004, pp. 462–469.
- [8] N. Fuhr, K. Grojohann, XIRQL: a query language for information retrieval in XML documents, in: W.B. Croft, D.J. Harper, D.H. Kraft, J. Zobel (Eds.), *Proc. of SIGIR*, ACM Press, New York, 2001, pp. 172–180.
- [9] W.H. Inmon, *Building the Data Warehouse*, John Wiley & Sons, New York, 1996.
- [10] R. Kimball, *The Data Warehouse Toolkit*, John Wiley & Sons, New York, 2002.
- [11] V. Lavrenko, W.B. Croft, Relevance-based language models, *Proc. of SIGIR*, ACM Press, New York, 2001, pp. 120–127.
- [12] D.M. Llidó, R. Berlanga, M.J. Aramburu, Extracting temporal references to assign document event-time periods, in: H.C. Mayr, J. Lazansky, G. Quirchmayr, P. Vogel (Eds.), *Proc. of DEXA*, Springer-Verlag, Berlin, 2001, pp. 62–71.
- [13] B.R. Moole, A probabilistic multidimensional data model and algebra for OLAP in decision support systems, *Proc. of IEEE SoutheastCon*, 2003.
- [14] Morgan Stanley Capital International Inc., <http://www.msci.com>.
- [15] B.-K. Park, H. Han, I.-Y. Song, XML-OLAP: a multidimensional analysis framework for XML warehouses, in: A.M. Tjoa, J. Trujillo (Eds.), *Proc. of DaWaK*, Springer-Verlag, Berlin, 2005, pp. 32–42.
- [16] T.B. Pedersen, C.S. Jensen, C.E. Dyreson, A foundation for capturing and querying complex multidimensional data, *Information Systems* 26 (5) (2001).
- [17] D. Pedersen, K. Riis, T.B. Pedersen, XML-extended OLAP querying, *Proc. of SSDBM*, IEEE Computer Society, Washington, 2002, pp. 195–206.
- [18] J.M. Pérez, R. Berlanga, M.J. Aramburu, A document model based on relevance modeling techniques for semi-structured information, in: F. Galindo, M. Takizawa, R. Traummüller (Eds.), *Proc. of DEXA*, Springer-Verlag, Berlin, 2004, pp. 318–327.
- [19] J.M. Pérez, T.B. Pedersen, R. Berlanga, M.J. Aramburu, IR and OLAP in XML document warehouses, in: D.E. Losada, J.M. Fernández-Luna (Eds.), *Proc. of ECIR*, Springer-Verlag, Berlin, 2005, pp. 536–539.
- [20] J.M. Ponte, W.B. Croft, A language modeling approach to information retrieval, *Research and Development in Information Retrieval*, ACM Press, 1998, pp. 275–281.
- [21] T. Priebe, G. Pernul, Towards integrative enterprise knowledge portals, *Proc. of CIKM*, ACM Press, New York, 2003, pp. 216–223.
- [22] I. Sanz, R. Berlanga, M.J. Aramburu, Gathering metadata from web-based repositories of historical publications, *Proc. of DEXA*, IEEE Comp. Society, 1998, pp. 473–478.
- [23] H. Schimdt, Probabilistic part-of-speech tagging using decision trees, *Proc. of Intl. Conf. on New Methods in Language Processing*, 1994.
- [24] G. Spofford, *MDX Solutions with Microsoft SQL Server Analysis Services*, John Wiley & Sons, New York, 2001.
- [25] W3C, Extensible Markup Language (XML) 1.0, <http://www.w3.org/TR/REC-xml>, 2004.
- [26] Xyleme, A dynamic warehouse for XML data of the Web, *IEEE Data Engineering Bulletin* 24 (2) (2001).



**Juan Manuel Pérez-Martínez** obtained a B.S. degree from the Universitat Jaume I (Spain) in 2000, where he is registered for a Ph.D. Currently he is associate lecturer at the same university. He is author of a number of communications in international conferences such as DEXA, ECIR, etc. His research interests are information retrieval, multidimensional databases, and web-based technologies. Contact him at [Juanma.Perez@lsi.uji.es](mailto:Juanma.Perez@lsi.uji.es).



**Rafael Berlanga-LLavori** is an associate professor in the Computer Science career at University Jaume I, Spain for 12 years. He received the B.S. degree from Universidad de Valencia in Physics, and the Ph.D. degree in Computer Science in 1996 from the same university. He is author of several articles in international journals, such as *Information Processing and Management*, *Concurrency: Practice and Experience*, *Applied Intelligence*, among others, and numerous communications in international conferences such as DEXA, ECIR, CIARP, etc. His current research interests are knowledge bases, information retrieval, and temporal reasoning. Contact him at [berlanga@lsi.uji.es](mailto:berlanga@lsi.uji.es).



**María José Aramburu-Cabo** is an associate professor in the Computer Science career at University Jaume I, Spain. She obtained the B.S degree from Universidad Politécnica de Valencia in Computer Science in 1991, and a Ph.D. from the School of Computer Science of the University of Birmingham (UK) in 1998. She is author of several articles in international journals, such as *Information Processing and Management*, *Concurrency: Practice and Experience*, *Applied Intelligence*, and numerous communications in international conferences such as DEXA, ECIR, etc. Her main research interests include document databases, and their applications. Contact her at [aramburu@icc.uji.es](mailto:aramburu@icc.uji.es).



**Torben Bach Pedersen** is an associate professor of Computer Science at Aalborg University, Denmark. His research interest includes multidimensional databases, OLAP, data warehousing, federated databases, data streams, and location-based services. He has published more than 60 scientific papers on these issues in journals and conferences such as *The VLDB Journal*, *Information Systems*, *IEEE Computer*, *VLDB*, *ICDE*, *SSDBM*, *SSTD*, *IDEAS*, *ACM-GIS*, *ECIR*, *Hypertext*, *DOLAP*, and *DaWaK*. He is a member of the Editorial Board of the *International Journal on Data Warehousing and Mining*, and has served on more than 30 program committees including *VLDB*, *ICDE*, *EDBT*, *SSDBM*, and *DaWaK*. Before joining Aalborg University, he worked in the software industry for more than six years. He received the Ph.D. and M.S. degrees in Computer Science from Aalborg University and Aarhus University, respectively. He is a member of the IEEE, the IEEE Computer Society, and the ACM. Contact him at [tbp@cs.aau.dk](mailto:tbp@cs.aau.dk).