CobWeb Multidimensional Model: Visualizing OLAP Query Results Using Tag-Cloud Operators

Omar Khrouf[1], Kaïs Khrouf[1], Jamel Feki[2]

[1] University of Sfax, MIR@CL Laboratory
Omar.khrouf@yahoo.fr, Khrouf.Kais@isecs.rnu.tn,

[2] University of Jeddah, FCIT, IS dept
jfeki@uj.edu.sa

Abstract. OLAP (On-Line Analytical Processing) systems provide decision makers with multidimensional analyses of large databases by generating an aggregated vision of data. Nowadays, these systems face growing non-numeric data. In this context, we propose in this paper a new generic multidimensional model called *CobWeb* dedicated to the OLAP of XML documents; it is based on the concept of facet. The *CobWeb* model aims to ease the expression of queries; also it offers an appropriate vision of the document warehouse. In this context and for manipulation purposes, we propose new visualization operators for OLAP query results by using the concept of Tag clouds as a means to help decision-makers to see the content in an efficient manner and then to focus on the knowledge in results.

Keywords: XML Documents, OLAP, *CobWeb* multidimensional model, Tag Clouds.

1 Introduction

Nowadays, documents contain pertinent data for the decision-making process and, therefore, their analysis can help decision makers to better understand the business processes of their organizations and take well founded decisions they cannot find by examining a conventional (numerical) data warehouse. However, in practice, most of the textual data are not included in decision-making systems or not considered during the decision-making process; this is due to several reasons as the lack of specific features for document warehouses. To alleviate this problem, it became necessary to integrate textual data into the decision-making information systems; this will help decision makers to identify hidden aspects and then improve their decisions.

In this context, several efforts have been invested; more accurately, most of them were interested in the exploitation of information stored in documents by using a set of facets [5] in order to model the documents according to several viewpoints issued from users and to improve the On-Line Analytical Processing (OLAP) of documents [11]. In this paper, we focus on the multidimensional modeling of documents and the OLAP analysis of textual data.

According to our review of the literature works related to document warehousing and OLAP of documents, the existing contributions could be classified into two main categories: (1) Works which adapted the classical multidimensional model (e.g., star, snowflake or constellation schemas) by enriching it with extensions specific for textual processing ([2] and [3] for data-centric documents; [6] for document-centric documents); and (2) Works which proposed specific models for the OLAP of documents, such as galaxy model [8] and diamond model [1].

The purpose of our work is to propose a new multidimensional model called *CobWeb*, as an extension of the galaxy model [9] dedicated to the OLAP of documents. The *CobWeb* model is based on the use of the concept of facets, these facets are standard. At the conceptual level a facet could be seen as a viewpoint since it groups a set of data describing of similar documents. At the external level (i.e., query level), a facet is considered as a means of expression for users' requirements. That is why we transform every facet into a multidimensional component of the *CobWeb* document warehouse model; more accurately a facet is transformed into a dimension as a potential axis of analyses. We have ensured that our *CobWeb* model differs from the existing models by a set of significant extensions such as the exclusion constraint between dimensions, the ability to define recursive parameters, duplicated dimension, and correlated dimensions.

From the other hand, and for performing analyses of the contents of documents, we have proposed a set of four operators for the visualization of results of OLAP queries by using the concept of Tag clouds in order to help decision-makers to better interpret the results of their queries.

We have organized this paper as follows. Section 2 presents the related works dealing with the OLAP of documents and highlights their main drawbacks. Section 3 describes the *CobWeb* multidimensional model focusing on its specifics. Then, we present, in Section 4, our new OLAP operators for document warehouses. Finally, Section 5 is reserved for the conclusion.

2 Related work

In this section, we focus on the related works dealing with the OLAP of documents. Among these works the authors in [10] proposed a framework for multidimensional analysis of *XML* documents called *XML-OLAP*. They defined a multidimensional expression language over *XML* cubes called *XML-MDX*. They also specify text mining operators for aggregating text constituting the measure data such as summarization, classification, and top keyword extraction.

Other studies suggest using the text mining techniques and information retrieval in order to aggregate textual data. In this sense, [9] developed two aggregation functions for the galaxy model: 1) *AVG_KW* summarizes a set of pseudo-average keywords replacing the initial set with a smaller and a general set; and 2) *TOP_KW* function that returns the k main keywords from a set of keywords. To do so, first the authors calculate the frequencies of terms by using the *Tf-Idf* function, secondly they select the first term among the k-most-frequent terms.

In [6] a textual dimension has a hierarchy that specifies the semantic between the terms of documents; the hierarchy enables a semantic navigation through two operators: *PULL-UP* and *PUSH-DOWN* in order to help users in the multi-semantic-level analysis of textual data. The authors define two aggregation measures adapted to textual data: Terms Frequency and the Inverted Index.

[9] and [6] have used analytical measurements based on the statistical method *Tf-Idf*. However, this method does not allow a consideration of the semantics content of documents. In order to integrate this aspect, the authors of [7] propose a new textual analysis measure for *CXT-Cube*, based on an adapted vector space model to represent textual data. In order to calculate the document terms' weights, [7] developed a relevance propagation technique on a concept hierarchy. Also, the authors provide an aggregation operator *Orank* (OLAP rank) that aggregated a set of documents by ranking them in a descending order using a vector space representation.

[1] present new aggregation operators that take into account the specificities of textual data from documents: *List_Concept* : returns a list of the most used concepts from a set of concepts in order to aggregate them into the corresponding cell of the multidimensional table; *G_Concept*: extracts the most used generic concepts; *S_Concept*: extracts the most used specific concepts; *Top_Concept*: groups the operators *List_Concept, G_Concept* and *S_Concept* to display the first concept of each of these operators.

We have learned from this study that most of works related to OLAP of documents proposed aggregate functions to deal with the textual content of documents. The result of these aggregate functions neglects, by eliminating, less frequent words which could be significant for the decision-maker analyses. To alleviate this limit, we propose, in this paper, a new visualization technique based on the Cloud of Tags format for the multidimensional result; this format highlights the most frequent concepts to the decision-maker without eliminating those are less frequent. In addition, the decision-maker will be able to examine less frequent concepts by hiding those are highly sized.

3 CobWeb multidimensional model

In order to integrate the concept of *facet* in our multidimensional model, we define a set of five facets namely *Content*, *Structural*, *Metadata*, *Keyword* and *Semantic* [4]. Each facet describes a useful aspect to the exploitation of documents according to a viewpoint in order to allow the user to see documents from multiple views and then to have a more targeted access to information as needed. In our approach, the five proposed facets should be: (1) *Standard*, i.e., independent of any specific domain of application, (2) *Document-Structure Free*, i.e., not limited to a set of predefined documents or to documents of the same structure, (3) Complete, i.e., cover all the information that describe XML documents and which are useful to the decision-maker to satisfy their OLAP needs, and finally (4) Automatically extracted from documents.

Based on these facets, we propose the *CobWeb* multidimensional model dedicated to the OLAP of documents in order to provide more opportunities for the expression of analytic queries and a vision more targeted of the data to decision-makers. To build this model, the main idea consists to transform every facet into a dimension since these facets may represent a means of expressing the users' viewpoints and therefore, cover the required data for their requirements. Furthermore, we have enriched the model with the *Document* dimension in order to link the information from different facets to their original documents.

In our work, we are inspired from the galaxy model [9] where a dimension could act as a subject of analysis (i.e., fact in the multidimensional terminology). This model connects several compatible dimensions by a node. However, the galaxy model is convenient for a collection of document which must be structurally homogeneous; this represents a restriction to decision-makers when they want to analyze data belonging to different collections. To alleviate this limit, our *CobWeb* model (Fig. 1) allows to group factual and textual data extracted from heterogeneous documents (Structure and content). *CobWeb* differs from the existing models by the following extensions:

- **Correlated Dimensions:** The conventional multidimensional operators (i.e., specific to numeric data warehouses) *Drill Down* and *Roll Up* allow detail/aggregate the results of OLAP analysis; they apply on parameters belonging to a same hierarchy. However, in the OLAP of documents, it would be interesting for example to move from the Content facet to the Concept facet or from Content to Keywords facet. In the classical multidimensional modeling, moving across dimensions is prohibited because of the absence of inter-dimensional relationships. In the *CobWeb*, we propose the concept of correlated dimensions, which allows a same query to move between dimensions (Fig. 1). The shifting from one dimension to another is accepted under condition: when we respect the direction of the dashed arrow linking dimensions. For example, it is possible to move from the *D_Content* dimension to the *D_Semantic* dimension, the reverse is not allowed.

- **Duplicated dimension:** It is a dimension that could be used more than once in the same OLAP analysis (multidimensional query). This type of dimension is rarely used in classical modeling (such as the star model); when used it requires high skilled user. In order to illustrate the duplicated dimension, suppose we want to perform analyses of documents published and sold within the same period (i.e., year for instance). Thus, the Date dimension must be used twice (once as a Publish date and once as a Sale date). Such a request requires a level of expertise raised for the decision-makers, who are generally not computer scientists.

In our works, we have proposed a specific dimension namely *Metadata* characterized by a set of additional parameters that can be used both in the same query, such as Author or Language. Graphically, a duplicated dimension is symbolized by the letter **D**. In the *CobWeb* model, we have only one duplicated dimension, called Metadata (Fig. 1).

- **Recursive Parameter:** In classical data warehouse schemas, parameters and hierarchies of dimensions are known in advance. However, in our work:

— The structure of a document may differ from one collection to another; it is hierarchically organized.

- We determine the semantic structure of documents by projecting the document on taxonomy (a hierarchy of concepts); this helps describing the textual content of documents. The number of concepts and levels varies from one taxonomy to another.

For these two dimensions define a new type of parameter called *recursive parameter*, since the documents and the taxonomy used in our model is represented in a hierarchical manner.

- The structural dimension helps us to move between levels (Content <Section <Subsection<Paragraph) using the conventional OLAP operators namely Roll Up and Drill Down.
- The semantic dimension allows the movement between concepts (e.g., from "Information System" to "Data Warehouse" to "Cube" and reversely).

Graphically, a recursive parameter is shown diagrammatically by a directed loop (Fig. 1).

- **Exclusion Constraint between Dimensions:** The exclusion constraint requires that a couple of dimensions cannot be used simultaneously in the same analysis. In *CobWeb*, the exclusion constraint concerns the Document and the Structural dimensions because an analysis must concern the documents or parts of the documents (title, section, etc.), but not both at the same time. Graphically, this exclusion constraint is denoted by a circle containing the letter **X** connected to the involved dimensions, such as: *D_Document* and *D_Structural* in Fig 1.

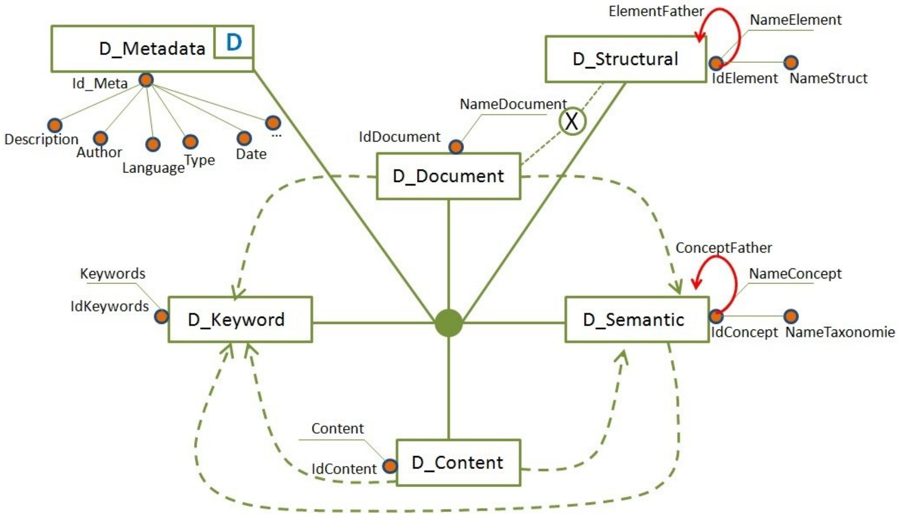


Fig. 1. CobWeb model.

4 New visualization operators for OLAP query results in document warehouses

For the OLAP of documents, the content of the multidimensional tables can be numerical (e.g., the number of documents) or textual (e.g., list of keywords, concepts). When the content of these multidimensional tables become very congested it complicates the task of the decision-maker. In the literature, most of works have proposed aggregation functions (e.g., *Top_Kewyord* [8] or *Top_Concept* [1]) in order to display the N most important keywords or concepts in the documentary collection. In our work, we propose to use the concept of tag clouds in order to represent the result of an OLAP query performed on a document warehouse. The tag cloud helps users (i.e., decision-makers) better see the result of a query and then better understand the result.

To do so, we propose a set of four operators for the visualization of results of OLAP queries as a means to help decision-makers focusing on knowledge incarnated in results. These operators are called *TableCloudTags, Filter_Tag, CellCloudTags* and *Agg_Tag*. They are detailed hereafter.

4.1 *TableCloudTags* Operator

This operator accepts the content of a multidimensional table resulted from an OLAP query on a document warehouse document and displays this content as a single cloud of words where every word has a font size proportional to its frequency in the table.

Definition. $Cube_{TC}$ = *TableCloudTags (Cube)* where:

Input

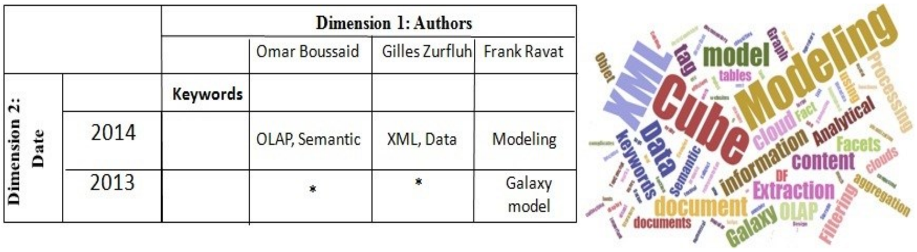
- *Cube* is the multidimensional table on which the *TableCloudTags* operator is applied,

Output

- $Cube_{TC}$ is the resulting cube composed of a single cell of Clouded tags.

Example

Query Q1: Analyze the keywords by Author and Date. Figure 2 depicts the multidimensional table which results from the execution of this OLAP query as a tag cloud. Note that keywords such as **XML**, **Cube** and **Modeling** have a higher frequency compared to remaining keywords. Therefore, the analyzed documents treat more particularly these concepts.



Dimension 2: Date		Dimension 1: Authors			
			Omar Boussaid	Gilles Zurfluh	Frank Ravat
		Keywords			
	2014		OLAP, Semantic	XML, Data	Modeling
	2013		*	*	Galaxy model

Modeling
Cube
XML
model
tag
Graph
tables
Processing
Data
cloud
Fact
Analytical
Facets
information
content
Keywords
Semantic
DF
Extraction
document
OLAP
documents
Galaxy
filtering
aggregation
clouds
Olap

Fig. 2. Result of query Q1

4.2 *Filter_Tag* Operator

The number of words generated by the tag cloud could be very important for a multidimensional query. For this reason, we propose to filter the tag clouds by:

- **User profile:** The user can specify his profile that describes his interests. In our work, the user profile is composed of a set of topics; each topic is associated with a set of keywords.
- ***Filtering according to a threshold****:* To reduce the number of words generated by the tag cloud, the user can set a threshold that represents the percentage of words to be displayed (for example, 30%). In this case, the tag cloud shows only the most significant words based on this percentage.
- ***The specification of values from data of dimensions:*** decision-makers can filter the tag cloud by choosing some values from dimension data of the multidimensional table.

Definition. $Cube_{FT}$ = *Filter_Tag (Cube, Type)* where:

Input

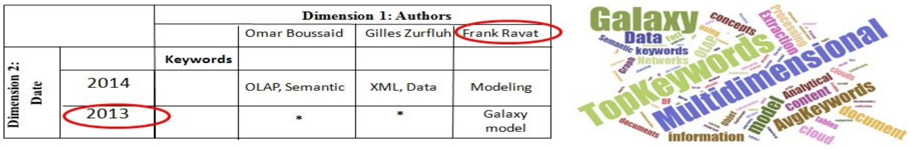
- Cube is the multidimensional table on which the *Filter_Tag* operator is applied,
- The selected type in order to filter the tag cloud

Output

- $Cube_{FT}$ is the resulting cube having the same number of cells as the input cube

Example

Figure 3 shows the tag cloud filtered by the *Author* dimension **(Frank Ravat)** and *Date* dimension **(2013)**.



Dimension 2: Date		Dimension 1: Authors			
			Omar Boussaid	Gilles Zurfluh	Frank Ravat
		Keywords			
	2014		OLAP, Semantic	XML, Data	Modeling
	2013		*	*	Galaxy model

Galaxy
concepts
Extraction
Data
Semantic
keywords
Networks
TopKeywords
Multidimensional
Analytical
content
AvgKeywords
model
document
information
documents
tables
cloud
DF

Fig. 3. Example of filtering the tag cloud by Data dimension

4.3 *CellCloudTags* Operator

In the multidimensional table, every word is displayed only once in the cell even if it was repeated (where is it repeated) many times. This representation imposes a restriction to decision-makers if they want to know the most analyzed keywords in every cell. To help, we propose a new OLAP operator: *CellCloudTags* in order to combine the OLAP and the tag cloud in order to represent the result of OLAP querying as a tag cloud in every cell of the multidimensional table. This representation allows us to know the most important data in each cell.

This operator accepts the content of a multidimensional table resulted from an OLAP query on a document warehouse document and displays this content as multiple clouds of words; in fact as many clouds as cells in the table. Naturally, every word in a cell has a font size proportional to its frequency in its cell.

Definition. *Cube_TagC= CellCloudTags (Cube, Cij)* where:

Input

- Cube is the multidimensional table, on which the Coup_Tag operator is applied,
- C_{ij} is the intersection of i^{th} line and the j^{th} column of the Cube

Output

- *Cube_TagC* represent the tag cloud in every cell of the multidimensional table.

Example

Figure 4 shows the result of query Q1 by using the operator *CellCloudTags*. We note that **Galaxy** has font size bigger than **XML**; therefore, the author Frank Ravat investigates more in the **Galaxy** than **XML**.

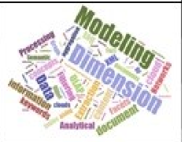
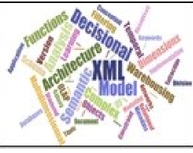
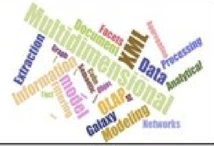
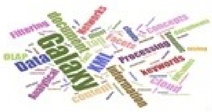
		Dimension 1 : Authors			
			Omar Boussaid	Gilles Zurfluh	Frank Ravat
Dimension 2 : Date		Keywords			
	2014				
	2013		*	*	

Fig. 4. Example of coupling multidimensional table and tag cloud

4.4 *Agg_Tags* Operator

This operator aggregates the textual data represented by our tag cloud in every cell of the multidimensional table in order to help the decision-maker to draw his attention on the most frequent data by dimensions axis. The aggregation function we are using calculates the number of occurrences of every word in each cell of the cube according to one among the two displayed dimensions. The relevance of each word in a dimension is proportional to its frequency in the set of words displayed in that dimension.

The navigational analysis through OLAP cubes massively uses aggregations through drilling operations such as the Roll Up and the Drill Down operations. In order to help the decision-maker to easily interact with an OLAP system, we have proposed a new OLAP operator: *Agg_Tag* in order to allow the aggregation of textual data in OLAP environments. The proposed aggregation operator has to combines different tag cloud by axis dimensions.

Definition. *Cube_Agg, = Agg_Tag (Cube_TagC, Method)* where:

Input

- *Cube_TagC* is the cube result of the operator *Coup_Tag*
- *Method ∈ {line, Column}* allows the user to aggregate by dimension 1(line) or dimension 2 (Column).

Output

- *Cube_TagC* represent the aggregation tag cloud by axis dimensions.

Example

Figure 5 shows the result of query Q1 by using the operator *Agg_Tag*.

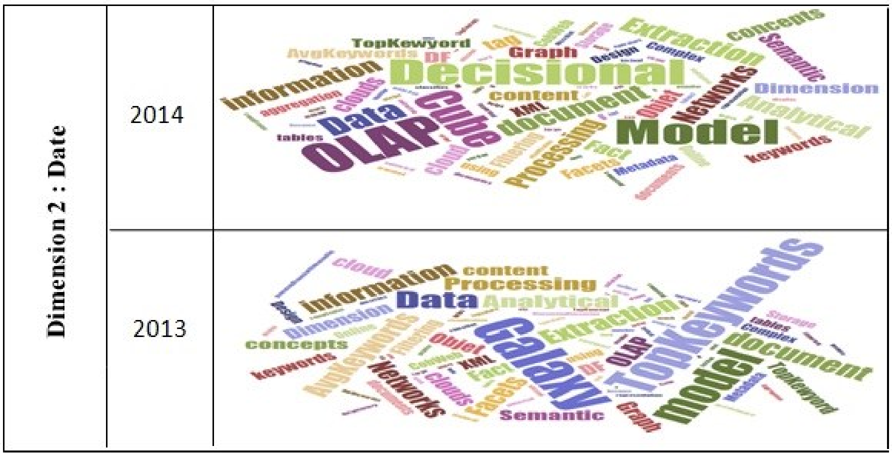


Fig. 5. Aggregation tag cloud by *Date* dimension

5 Conclusion

In this paper, we presented a framework for a multidimensional OLAP text analysis. We proposed a new generic multidimensional model dedicated to the On-Line Analytical Processing (OLAP) of XML documents, called *CobWeb* model. This model differs from existing models by the following extensions: the exclusion constraint, the recursive parameters, the duplicate dimension and the correlated dimensions. Thus, we proposed a new visualizing OLAP query results based on tag clouds in order to facilitate the interpretation of the results of the multidimensional analyses on the textual data. In perspective, we plan to conduct experiments to measure the quality of the result extracted by our OLAP operators. Finally, we intend to introduce the collaborative aspect which allows the sharing of OLAP analyses between users working in the same organization.

References

1. Azabou, M., Khrouf, K., Feki, J., Vallès, N., Soulé-Dupuy, C.: Diamond multidimensional model and aggregation operators for document OLAP. IEEE Ninth International Conference on Research Challenges in Information Science 2015, Athens, Greece, p. 363-373.
2. Feki, J., Ben Messaoud., I., Zurfluh, G.: Building an XML Document Warehouse. Journal of Decision Systems (JDS), Taylor & Francis, vol. 22, number 2, p. 122-148.
3. Hachaichi, Y., Feki, J.: An Automatic Method for the Design of Multidimensional Schemas from Object Oriented Databases. International Journal of Information Technology and Decision Making, vol. 12, number 12, p. 1223-1259.
4. Khrouf, O., Khrouf, K., Altalhi, A., Feki, J.: CobWeb Multidimensional Model: Filtering Documents using Semantic Structures and OLAP. ICIW 2015: The Tenth International Conference on Internet and Web Applications and Services. IARIA, 2015. ISBN: 978-1-61208-412-1.
5. Kumar, S., Morstatter, F., Marshall, G., Liu, H., Nambiar, U. Navigating Information Facets on Twitter (NIF-T). Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China, p. 1548-1551.
6. Lin, C. X., Ding, B., Han, J., Zhu, F., Zhao, B.: Text cube: Computing in measures for multidimensional text database analysis. Eighth IEEE International Conference on Data Mining 54, p. 905–910.
7. Oukid, L., Asfari, O., Bentayeb, F., Benblidia, N., Boussaid, O., CXT-cube: contextual text cube model and aggregation operator for text OLAP.
8. Ravat, F., Teste, O., Tournier, R., Zurfluh , G.: Designing and Implementing OLAP Systems from XML Documents. Proc. Annals of Information Systems, Springer, Special issue New Trends in Data Warehousing and Data Analysis, vol. 3, p. 1-21.
9. Ravat, F., Teste, O., Tournier, R., Zurfluh , G.: Finding an Application-Appropriate Model for XML Data Warehouses. Information Systems, Elsevier, Vol. 36 N. 6, p. 662-687.
10. Park, B.-K., Han, H., Song, I.-Y.: XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses. In 7th International Conference on Data Warehousing and Knowledge Discovery, Volume 3589 of LNCS, p. 32–42. Springer.
11. Zhang, D., Zhai, C., Han J.: Topic cube: Topic modeling for olap on multidimensional text databases. SDM ’09: Proceedings of the 2009 SIAM International Conference on Data Mining, Sparks, NV, USA, p. 1124–1135.