CobWeb Multidimensional Model: Filtering Documents using Semantic Structures and OLAP

Omar Khrouf, Kaïs Khrouf
University of Sfax, MIR@CL
Laboratory, Tunisia
Omar.khrouf@yahoo.fr,
Khrouf.Kais@isecs.rnu.tn

Abdulrahman Altalhi
Faculty of Computing and IT,
King Abdulaziz University,
Jeddah, Saudi Arabia
ahaltalhi@kau.edu.sa

Jamel Feki
Faculty of Computing and IT,
University of Jeddah,
Jeddah, Saudi Arabia
Jamel.Feki@gmail.com

***Abstract*—Today, the documents constitute a capitalization of knowledge in the Information Systems of companies. For the decision-makers, analyzing the contents of documents represents a real challenge. This paper proposes an approach based on the *CobWeb* model to filter semantic structures in order to find documents relevant to the decision-makers' needs. In order to validate our approach, we have developed a GUI for the multidimensional queries and we have applied the Online Analytical Processing (OLAP) analysis on 250 documents taken from the academic domain.**

Keywords-XML documents; standard facet; OLAP; multidimensional model.

I. INTRODUCTION

The information systems of companies accumulate, over time, an important volume of data. With the web applications, the users can improve the internal communication within the company by creating, sharing, and modifying real-time work files. The information sharing and the professional work are essential for communication and business productivity. Faced with the rapid development of data (particularly in Web applications), the decision-making process has become an essential activity and an important research area, which requires the implementation of efficient systems called Decision Support Systems (DSS). In addition to the classical DSS systems, which handle numeric data, several studies have been interested in the exploitation of documentary information in order to extract semantic knowledge, for example, the multi-representation of documents using a set of "Facets" [8], or the OLAP of documents [13].

For the multi-representation of documents, some authors, such as Hernandez and al. [8] and Charhad and al. [4] have proposed to use a set of facets in order to describe the useful aspects of documents. These facets take into account not only the semantic aspect, but also other factors related to the exploitation context of documents in order to better satisfy the users' needs. However, the various proposed facets vary according to the application field. Therefore, it would be interesting to define the standard facets, which enable the representation of documents in any research area.

For the OLAP of documents, two categories of works can be distinguished: (1) Those which have adopted the classical multidimensional model, i.e., the star, the snowflake and the constellation models by enriching them with extensions for textual processing ([6] and [7] for data-centric documents; [15] for document-centric documents); and (2) Those which have proposed specific models for the OLAP of documents, such as *galaxy model* [13] and *diamond model* [1]. However, these studies did not treat the heterogeneity of structures and hence, require the definition, in advance, of parameters and hierarchies.

In order to give more flexibility to the user in OLAP analysis tasks, we have proposed a multidimensional model called "*CobWeb model*", as an extension of the galaxy model dedicated to the OLAP of documents based on standard facets [9]. Each facet includes a set of data and is considered as a means of expression for the user's needs. For this reason, we have transformed every facet into a dimension. In multidimensional modeling, each dimension has a structure composed of a set of attributes called parameters, arranged hierarchically from the finest to the highest granularity (e.g., the Time dimension is composed of: Day < Week < Month < Quarter < Semester < Year). The dimension can be considered as an analysis axis; a parameter represents an analysis level and may be associated with one or several descriptive attributes, commonly called weak attributes. However, the integration of facets into an OLAP model creates a set of new problems, for which the classical models of the literature are unable to solve. Such problems arise from the multiple use of the same dimension within the same analysis, and from the concept of recursivity for a parameter of a given hierarchy, etc. To overcome these problems, we have proposed a set of extensions in the *CobWeb* model such as the exclusion constraint between two dimensions, which doesn't allow using these two dimensions in the same analysis. The recursive parameters are used when the hierarchy parameters are not known in advance. The duplicated dimension allows the use of the same dimension twice in the same analysis, whereas the correlated dimension enables the movement between dimensions in the same analysis.

This paper introduces the *CobWeb* model concepts and presents our approach of document filtering by using Semantic Structures. It is organized as follows. Section 2 presents the related works dealing with the representation and exploitation of facets of documents and the OLAP of documentary information. Then, we define in Section 3 the set of five proposed standard facets of documents. Section 4 describes the *CobWeb* multidimensional model focusing on its specificities. Section 5 presents the filtering of documents

ISBN: 978-1-61208-412-1

and the OLAP querying. Finally, Section 6 is reserved for the conclusion.

II. Related works

In this section, we first overview the related works dealing with the multidimensional modeling of documents. Then, we examine the major works dealing with the representation and the exploitation of facets extracted from documents.

For the multidimensional modeling of documents, most works have adopted the three proposed models in the literature for the factual data (star model, snowflake model and constellation model [10]) and have suggested some approaches or functions for the analysis of textual content.

Tseng and al. [14] have used the star schema in order to analyze documents. This schema distinguishes between three types of dimensions: metadata (describing the document, e.g., author, language), ordinary (an ordinary dimension contains keywords extracted from the document), and category (external data for the document description as issued from Wordnet). However, it is limited to a simple count of general documents (e-mails, articles, Web pages, etc.) according to dimensions.

Boussaid and al. [3] have proposed a modeling in snowflake of multidimensional XML data with data mining methods. These studies allow the analysis of complex data, but are not adapted for the analysis of textual data from XML documents.

Azabou and al. [1] have proposed a *diamond model*, which is the star model enriched with a central dimension that attempts to represent the semantics of the document. The parameters of this semantic dimension are linked to parameters of other dimensions. The main disadvantage of this work is that it proposes a model made by a collection of documents with the same structure. Ravat and al. [13] have proposed a multidimensional model, called *Galaxy*, which is adapted to the analysis of XML documents. A galaxy schema is uniquely based on the dimension concept; it connects several dimensions by nodes instead of facts. A connecting node denotes compatible dimensions for analysis. However, this work does not take into account the heterogeneity of document structures.

Zhang and al. [15] have proposed a new model called *Topic Cube,* based on the star schema which extends the traditional data cube by integrating a hierarchy of topics as an analytical dimension. It is a new cube model using a topic dimension and a text content measure which uses parameters of a probabilistic model. However, *Topic Cube* supports only a predefined set of themes.

We notice that the studies dealing with the OLAP of documents provide the analysis of the documents having the same or similar structures.

For the representation of documents, the concept of facet has been used in several domains and with different types of documents.

For tweets, Kumar and al. [11] have proposed a navigation system by facet called Navigating Information Facets on Twitter *(NIF-T)* based on three facets: the Geo Facet showing the location of tweets in a map. Subject facet is a word showing the different thematic exchanges by the tweets. Time facet presents the number of tweets in a given date.

For the video documents, Charhad and al. [4] have proposed to widen the Extended Model for Image Representation and Retrieval *(EMIR²)* created by Mechkour [12] in order to include audiovisual documents. They have added two facets: the temporal facet and the event facet. These two facets characterize the dynamic aspect, which is specific for this type of document. This new model allows the synthetic and integrated consideration of information about the image, text and sound elements.

For textual documents, Hernandez and al. [8] have proposed a model based on a multi-facet representation of documents in order to associate several facets into the same document. They have defined two types of facets: the first one represents the semantics of the contents and the second one includes parameters aiming to improve the research results of documents, such as the description of the educational theories, the description by metadata, etc.

As a conclusion, we notice that some of the studies which have used various facets vary depending on the application domain. However, for other approaches, the facets are fixed for a specific application domain.

The purpose of this paper is to integrate the notion of facet in the OLAP model because it is interesting to represent documents from several points of view. Then we propose an approach based on the *CobWeb* model to filter semantic structures in order to determine the documentary information for the user’s needs.

III. Standards facets of documents

To the best of our knowledge, the concept of facet has not been addressed in the decision domain. In order to provide a facet-based OLAP model, we define a set of five facets to represent one or many documents according to a given viewpoint. These facets must be standard, i.e., independent of any specific domain of application and must give the user the ability to consider the same document or set of documents from multiple views (Metadata, Keyword, etc), so that he can have a more targeted access to information as needed [9].

- The *Metadata* Facet: this facet aims to provide the users with a structured collection of the data describing a document (such as: title, rights, format, etc.). In our work, we use the metadata defined by the Dublin Core [5].
- The *Keyword* Facet: this facet constitutes a set of the most important keywords describing the content of the document. These keywords can be determined, by using the indexing techniques of information retrieval, or they come from the document itself when they exist explicitly.
- The *Content* Facet: this facet aims to present the information contained in the document (image, text,

etc.) by removing everything about the comments, structure, etc.

- The *Semantic* Facet: this facet describes the semantics of the content of the document. It is used in the classification of all or parts of the documents in order to facilitate the retrieval /analysis of these documents. For the determination of this semantics, we have relied on the work in [2] which defines a method for the determination of a semantic structure for a given document.
- The *Structural* Facet: this facet is a viewpoint of the structure of a document. It aims to focus on parts of the document (section, subsection, etc.) and not the whole document.

Based on the previously defined facets, we present, in the following sections, the *CobWeb* multidimensional model devoted to the OLAP of documents. Then, we present an approach to filter documents by using semantic structures and OLAP querying.

IV. COBWEB MULTIDIMENSIONAL MODEL

In this section, we present the *CobWeb* multidimensional model (Fig. 1), which is an extension of the galaxy model based on standard facets in order to provide more opportunities for the expression of analytic queries and a more targeted vision of the data to decision makers. To build this model, the main idea consists in transforming every facet into a dimension since these facets may represent a means of expressing the users' viewpoints and therefore, describe their requirements. Besides, we have added the dimension *Document* in order to link the information from different facets to their documents. *CobWeb* differs from the existing models by the following extensions:

- **Duplicated Dimension:** The classical multidimensional modeling does not allow using the same dimension twice in the same analysis. Let us suppose that we want to analyze the documents by two parameters belonging to the same Metadata dimension (namely, *Date* and *Editor*). This type of query is not possible. In order to give more flexibility to the user in the task of OLAP analysis, we propose the duplicated dimension, which can participate many times in the same analysis. Graphically, a duplicated dimension is symbolized by the letter **D** in the concerned dimension. In the *CobWeb* model, we have only one duplicated dimension, called *Metadata* (Fig. 1).
- **Recursive Parameter:** In the classical schema of data warehouses, the parameters and dimension hierarchies are known in advance. However, in our work:
    - ○ The structure of documents may differ from one collection to another.
    - ○ The semantic structure of documents is determined from taxonomies (hierarchical representation of the concepts) and helps describe the textual content of documents. Specifically, the concepts of taxonomies will be assigned to different parts of the documents. Therefore, the number of concepts and levels varies from a semantic structure to another. For the representation of these two dimensions, we will use a new type of parameters, called recursive parameter, since the documents and the taxonomies are represented in a hierarchical manner.
    - ○ The structural dimension helps us to move from one level to another (Content →Section →SubSection →Paragraph) using the conventional OLAP operators namely *RollUp* and *Drill Down*.
    - ○ The semantic dimension allows the movement between concepts (Information System →Data Warehouse →Cube, etc.).

  Graphically, a recursive parameter is schematized by a directed loop (Fig. 1).
- **Correlated Dimensions:** In the classical multidimensional modeling, the movements between the dimensions cannot be achieved because of the absence of inter-dimensional relationships. To solve this problem, we propose the concept of correlated dimensions which allows, for the same query, to move between dimensions. Graphically, the correlation that can be possible between the dimensions of our multidimensional model is represented by dashed arrows between dimensions. The transition from one dimension to another is accepted when we respect the direction of the arrow. For example, it is possible to move from the Content dimension to the Semantic dimension.
- **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 *Content* and the *Structural* dimensions because an analysis must concern the content or parts of the documents (title, section, paragraph, 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: *Document* and *Structural* in Figure 1.

V. FILTERING AND OLAP QUERING

In this section, we describe our approach of filtering documents using Semantic Structures and OLAP querying in order to find the documentary information relevant to the decision-makers' needs according to several analysis axes, as shown in Figure 2.

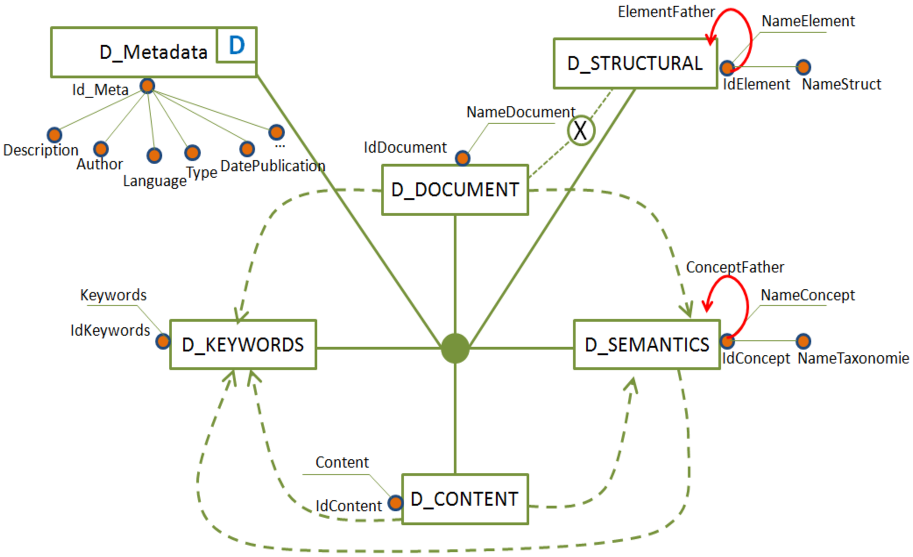


Figure 1. *CobWeb* multidimensional model.

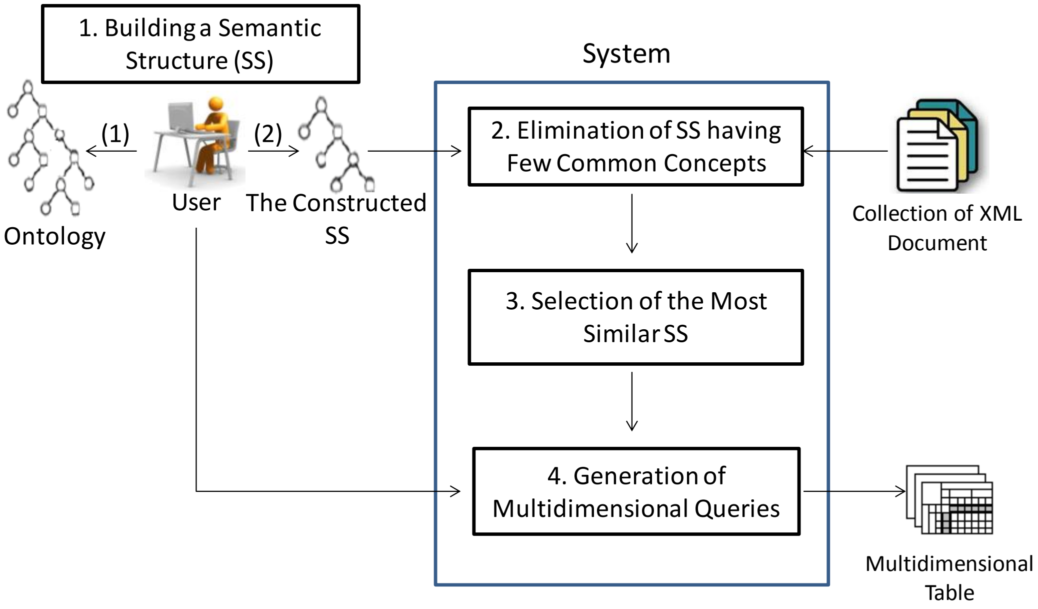

1. Building a Semantic Structure (SS)
System
(1)
(2)
Ontology
User
The Constructed SS
2. Elimination of SS having Few Common Concepts
Collection of XML Document
3. Selection of the Most Similar SS
4. Generation of Multidimensional Queries
Multidimensional Table

Figure 2. The proposed process of querying

.
This approach is based on four fundamental stages:

- The first step is building a semantic structure from a selected ontology. The user defines the closest concepts to his needs and puts them together into a semantic structure according to a set of predefined rules.

- The second step consists in a first filtering of the documents to be analyzed by the elimination of the semantic structures of documents having few common concepts with the built semantic structure defined by the user.

- The third step is a second filtering of documents in order to keep those having the most similar semantic structures compared to the semantic structure defined by the user.

- The last step in our approach is the generation of the multidimensional queries and the result will be displayed as a multidimensional table.

We note that we have automatically generated, in our previous works, a semantic structure for every XML document; this semantic structure is superposed on its logical structure [2] (Fig. 3).

The purpose behind the use of semantic structures, in this phase of querying, is to keep only the relevant documents for the user's need. In what follows, we explain the different steps of our approach.

A. *Step 1:Building a Semantic Structure (SS)*

In order to build his semantic structure, the user chooses, through a web application, a semantic resource to select a set of concepts depending on his needs and organizes them in a hierarchical way. This web application allows communicating and exchanging the semantic structures between systems or applications in order to determine the documentary information for the user's needs. The user will be assisted by the system that displays error messages for the incorrect manipulations and it suggests one or more solutions (Fig. 4).

A semantic resource (ontology, taxonomy, thesaurus, etc.) serves to represent the semantics of a given domain in a generic and reusable way in order to share knowledge and data.

To build his semantic structure, the user must respect the following rules:

- Rule 1: No reverse hierarchical order between concepts. Example: the ontology of Figure 4 shows that the *OLAP* is the father concept of the *Dimension*, in the semantic structure built by the user which it is prohibited to represent the *Dimension* as the father concept of the *OLAP*.
- Rule 2: The conceptual father of the concepts selected by the user represents the common ancestor of these concepts in ontology. Example: Figure 4 shows that the user has selected the *OLAP* and *Dimension* concepts. The conceptual father attributed to these concepts is *Data warehouse* because, in ontology, it represents the common ancestor of selected concepts.

B. *Step 2: Elimination of SS with Few Common Concepts*

Once the semantic structure is built by the user, we compare this structure with the semantic structures of documents, in order to keep the pertinent structures, i.e., eliminate the structures having few common concepts with the semantic structure built by the user. For this reason, we propose the measure of similarity (1).

$$\mathrm{Sim}_{(SSC,\, SSU)} = \frac{|C_{SSC}|}{|C_{SSU}|} \qquad (1)$$

$|C_{SSC}|$: Number of common concepts between the semantic structure of the user and the semantic structures of documents.

$|C_{SSU}|$: Number of concepts in the semantic structure of the user.

$\mathrm{Sim}_{(SSC,\ SSU)}$: The similarity degree between the two semantic structures.

We define a threshold for selecting structures. This threshold is determined by experiments and may be modified by the user according to his need.

C. *Step 3: Selection of the most Similar SS*

In our work, the order of concepts (father-son relation) is very important so, we propose to start comparing branches. A branch is a path composed of all the concepts between the root and the leaf of the semantic structure. Example: the branches of the semantic structure built by the user (Fig. 4) are:

Branch 1: Information System → Graph.

Branch 2: Information System → Data Base.

Branch 3: Information System → Data Warehouse → Dimension.

Branch 4: Information System → Data Warehouse → OLAP.

The measure of similarity (2) allows comparing two branches of both semantic structures.

$$\mathrm{SimB}_{(SSC,\, SSD)} = \left(\frac{|CA_{(SSC,\, SSD)}|}{\left(|CA_C| + |CA_D|\right)/2} + \frac{|AA_{(SSC,\, SSD)}|}{|AA_C|} \right) \Big/ 2 \qquad (2)$$

$|CA_{(SSC,\ SSD)}|$: Number of common concepts between the two mapped branches.

$|CA_C|$: Number of concepts in the branch of the built semantic structure.

$|CA_D|$: Number of concepts in the branch for the semantic structure of the document.

$|AA_C|$: Number of arches in the branch of the built semantic structure.

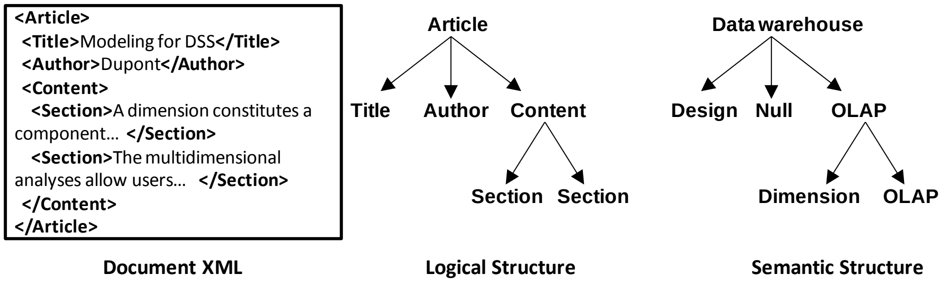
$|AA_{(SSC,\ SSD)}|$: Number of common arches between the two mapped branches.

Table I presents an example of comparison between two branches of two different documents with a branch of the semantic structure built by the user (Information System (IS) → Data warehouse (DW) → Dimension (DIM)).

TABLE I. A COMPARISON TABLE BETWEEN TWO BRANCHES OF TWO SEMANTIC STRUCTURES

Branche of SSC	Branche of SSD	Degree of similarity $\mathbf{SimB}_{(SSC,\, SSD)}$
IS **DW** **DIM**	**IS** **X** **DW** **DIM**	✓3 common concepts(**IS**, **DW**, **DIM**) ✓One arch aligned: **DW** ⟶ **DIM** $\mathrm{SimB}_{(SSC,\, SSD)} = \left(\frac{3}{(3+4)/2} + \frac{1}{2}\right)/2$ $= (0.85 + 0{,}5)/2$ $= 0.67$
	IS **DIM**	✓2 common concepts(**IS**, **DW**) ✓ 0 arch aligned $\mathrm{SimB}_{(SSC,\, SSD)} = \left(\frac{2}{(3+2)/2} + \frac{0}{2}\right)/2$ $= 0.8 + 0\ /2$ $= 0.4$

ISBN: 978-1-61208-412-1



```
<Article>
  <Title>Modeling for DSS</Title>
  <Author>Dupont</Author>
  <Content>
    <Section>A dimension constitutes a component… </Section>
    <Section>The multidimensional analyses allow users… </Section>
  </Content>
</Article>
```

Figure 3. Logical and semantic structure of XML document.

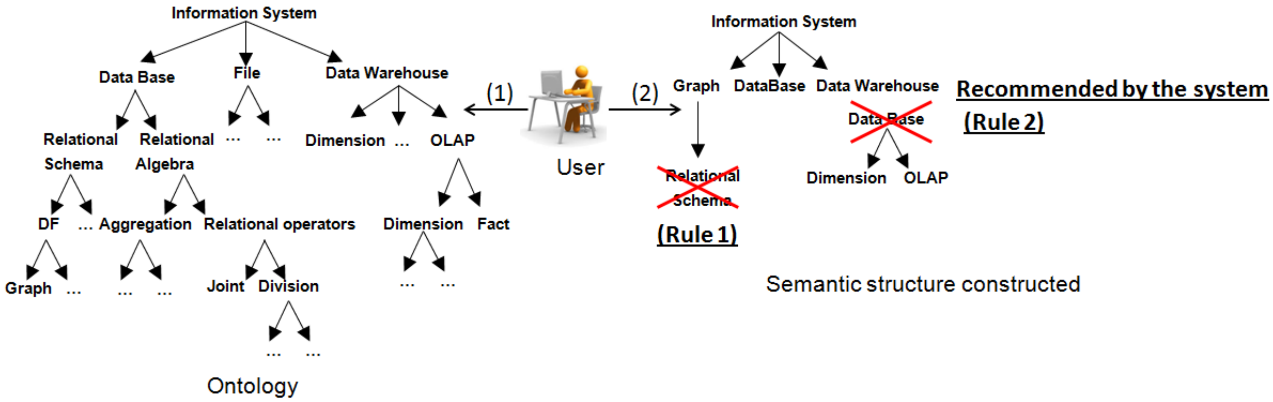


Figure 4. Example of building a semantic structure.

At this level, we need to calculate the weight sum of the different branches. The documents having a similarity degree above a threshold (fixed by experiment and may be modified by the user) will be selected for the OLAP Query.

D. Step 4: Generation of Multidimensional Queries

For the multidimensional querying, the user should specify his query by indicating the fact and its measures and the various dimensions. Then, the system automatically generates the needed queries.

We have developed a GUI for the multidimensional querying. In the left part of the interface, the user specifies his request by indicating the dimensions and the fact. The right part is devoted to the results of the query as a multidimensional table. To validate our work, we test and evaluate our approach on 250 documents taken from the academic domain.

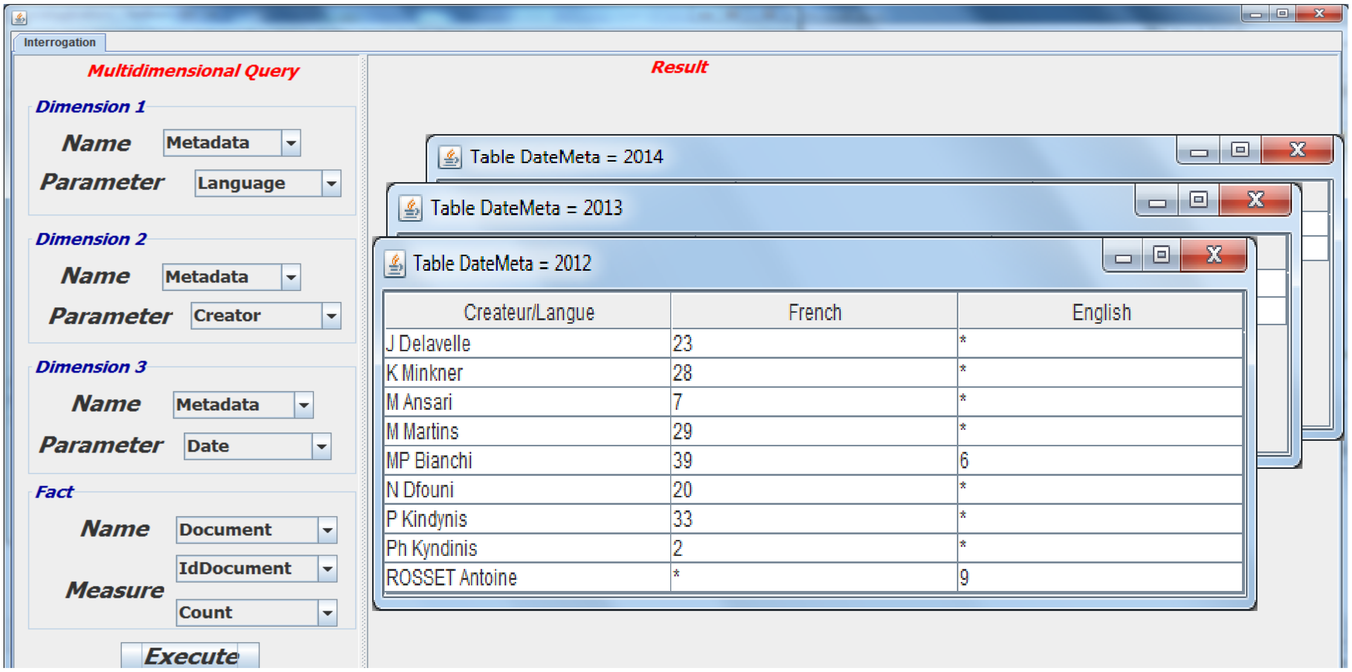
Figure 5 shows the results of the previous query in a multidimensional table, where the columns and the lines represent the first two dimensions (*Author* and *Language*), and where the plans represent the third dimension (*Date*).

The measures are placed in the intersection of a line and a column for a given plan. The symbol * indicates that there is no value for the measure. In this experiment, we observe that most documents are written in French (181 documents in 2012). So we can note that the majority of authors are francophone.

VI. CONCLUSION

In this paper, we have proposed an approach based on the *CobWeb* model to filter the documents using Semantic Structures in order to determine the documentary information for the user's needs. In addition, we have developed a GUI for the multidimensional querying.

The main limitation of our approach is that the user builds his semantic structure by using a single semantic resource; it would be interesting to offer more opportunities to the user in order to build his semantic structure from several resources and not just to one. We also intend to propose new OLAP operators that take into consideration the specificities of the *CobWeb* model, for example an operator for the correlation of dimensions. These operators will facilitate the interpretation of the results of the multidimensional analyses. In the long run, we plan to introduce the personalized OLAP analysis which takes into account the needs and skills of the users, based on their profiles.



Dimension 1
Name: Metadata
Parameter: Language

Dimension 2
Name: Metadata
Parameter: Creator

Dimension 3
Name: Metadata
Parameter: Date

Fact
Name: Document
Measure: IdDocument
Count

Table DateMeta = 2014

Table DateMeta = 2013

Table DateMeta = 2012

Createur/Langue	French	English
J Delavelle	23	*
K Minkner	28	*
M Ansari	7	*
M Martins	29	*
MP Bianchi	39	6
N Dfouni	20	*
P Kindynis	33	*
Ph Kyndinis	2	*
ROSSET Antoine	*	9

Figure 5. Graphical multidimensional querying.

REFERENCES

[1] M. Azabou, K. Khrouf, J. Feki, C. Soulé-Dupuy, and N. Vallès, "A Novel Multidimensional Model for the OLAP on documents: Modeling, Generation and Implementation," International Conference on Model & Data Engineering, Larnaca, Cyprus, September 2014, pp. 258–272.

[2] S. Ben Mefteh, K. Khrouf, J. Feki, and C. Soulé-Dupuy, "Semantic Structure for XML Documents: Structuring and pruning," Journal of Information Organization, vol. 3, issue 1, March 2013, pp. 37-46.

[3] O. Boussaid, R. Ben Messaoud, R. Choquet, and S. Anthoard, "X-Warehousing : An XML-Based Approach for Warehousing Complex," Proc. The 10 th East-European Conference on Advances in Databases and Information Systems (ADBIS 06), vol. 4152, September 2006, pp. 39-54.

[4] M. Charhad and G. Quénot, "Semantic Video Content Indexing and Retrieval using Conceptual Graphs," Proc. IEEE Information and Communication Technologies: From Theory to Applications (ICTTA 04), IEEE Press, April 2004, pp. 399-400, doi: 10.1109/ICTTA.2004.1307800.

[5] Dublin Core Metadata Initiative (DCMI), Dublin Core Metadata Element Set, Version 1.1, ISO Standard 15836, 2007. http://dublincore.org/documents/dces/.

[6] J. Feki, I. Ben Messaoud, and G. Zurfluh, "Building an XML Document Warehouse," Journal of Decision Systems (JDS). Ed. Taylor & Francis, vol. 22, issue 2, April 2013, pp. 122-148, doi: 10.1080/12460125.2013.780322.

[7] Y. Hachaichi and J. Feki, "An Automatic Method for the Design of Multidimensional Schemas from Object Oriented Databases," International Journal of Information Technology and Decision Making, vol. 12, issue 12, November 2013, pp. 1223-1259, doi : 10.1142/S0219622013500351.

[8] N. Hernandez, J. Mothe, B. Ralalason, B. Ramamonjisoa, and P. Stolf, "A Model to Represent the Facets of Learning Objects," Interdisciplinary Journal of E-Learning and Learning Objects, Information Science Institute, Santa Rosa - USA, vol. 4, January 2008, pp. 65-82.

[9] O. Khrouf, K. Khrouf, and J. Feki, "CobWeb Multidimensional Model: From Modeling to Querying," International Conference on Model & Data Engineering, Larnaca, Cyprus, September 2014, pp. 273–280.

[10] R. Kimball and M. Ross, The Data Warehouse Toolki: The Definitive Guide to Dimensional Modeling, 3rd ed., John Wiley & Sons, New York, July 2013.

[11] S. Kumar, F. Morstatter, G. Marshall, H. Liu, and U. Nambiar, "Navigating Information Facets on Twitter (NIF-T)," Proc. The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 12), August 2012, pp. 1548-1551.

[12] M. Mechkour, "A multifacet formal image model for information retrieval," Proc. The Final WorkShop on Multimedia Information Retrieval (MIRO 95), September 1995, pp. 18-20.

[13] F. Ravat, O. Teste, R. Tournier, and G. Zurfluh, "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, November 2008, pp. 1-21, doi: 10.1007/978-0-387-87431-9_15.

[14] F. S. C. Tseng and A. Y. Chou, "The concept of document warehousing for multidimensional modeling of textual-based business intelligence," Journal Decision Support System (DSS), vol. 42, issue 2, November 2006, pp. 727-744.

[15] D. Zhang, C. Zhai, J. Han, A. Srivastava, and N. Oza, "Topic modeling for OLAP on multidimensional text databases: topic cube and its applications," Journal Statistical Analysis and Data Mining, vol. 2, issue 5-6, December 2009, pp. 378-395.

ISBN: 978-1-61208-412-1