ISSN 2347 - 3983
Volume 7, No. 11 November 2019
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter387112019.pdf
https://doi.org/10.30534/ijeter/2019/387112019

Multidimensional Analysis of XML Documents: Modeling and Implementation

Kais Khrouf[1], Tarek Lefi[2]
[1]College of Computer and Information Sciences, Jouf University, Saudi Arabia, kmkhrouf@ju.edu.sa
[2]National School of Electronics and Telecommunications, University of Sfax, Tunisia, tarik.lefi@gmail.com

ABSTRACT

In order to access, visualize, aggregate and analyze their data, the organizations generally use the techniques of data warehouses and OLAP (On-Line Analytical Processing). We distinguish two categories of data: structured (such: databases), and unstructured (such: Documents). In order to store and manage these both kinds of data, organizations exploit increasingly the advantages of XML (eXtensible Markup Language); it constitutes an important source for decision-makers to help them to interpret and control the evolution of their business processes. In this paper, we propose a new multidimensional model of semi-structured data (XML documents), based on complementary dimensions (Content, Metadata, Structure, Semantics…) and we integrate also the concept of Tag Cloud for visualizing multidimensional tables, in order to help the decision-makers to better analyze textual data in the result of the OLAP queries.

Key words: Multidimensional Analysis, OLAP (On-Line Analytical Processing), XML documents, Tag Cloud.

1. INTRODUCTION

Nowadays, documents should be integrated into the decision support system since the volume of unstructured data grows faster than the volume of structured. The On-Line Analytical Processing (OLAP) technology provides a flexible representation of data in different granularities; we consider that it can be a useful technology to exploit unstructured data. The integration of unstructured data in OLAP systems and aggregating them in order to improve the decision-making constitute a challenge for Business Intelligence systems.
XML (eXtensible Markup Language) allows the exchange the data on the Web. The XML documents are text-oriented XML documents (Example: company reports, Scientific papers) and are less structured than data-oriented XML documents (Example: Documents generated from databases). It is important than to define new visualization techniques for unstructured data in OLAP systems and to improve the traditional existent data models.
In this context, we integrate the concept of Tag Cloud in multidimensional tables in order to help decision makers to visualize and interpret easily the results of their queries. A Tag Cloud is a visualization technique for displaying in a graphical representation the multidimensional table. It shows textual data (terms) as big as their frequency is high.
The objective of this work is to propose a new multidimensional model of XML documents extended by the possibility to integrate the concept of Tag Cloud in the result of an OLAP query (Multidimensional tables). For the concept of Tag Cloud, every term has a font size proportional to its frequency of occurrence in the OLAP query result.
This paper is organized as follows. Section 2 presents the related work dealing with the multidimensional modeling and the OLAP of XML documents. Section 3, describes the new multidimensional model we propose. Then, we present the software prototype that we have developed for the OLAP of documents. Finally, we provide the conclusion in Section 5.

2. LITERATURE REVIEW

In this section, we study the related works dealing with the multidimensional modeling and the OLAP of documents.
For the multidimensional modeling of documents, several works have been proposed, based generally on the star schema.
Reference [12] proposes a *Topic Cube* model. It is a star schema extended by a new dimension *Topics* generated from domain ontology according to the analysts' preferences. The authors of [11] use the star schema and propose three types of dimensions for modeling documents. The *Ordinary Dimension* constitutes a list of terms extracted from the document. The *Metadata Dimension* presents the information about the documents (Language, Author, etc.). The *Category Dimension* describes external data (according to users' viewpoints) concerning the document description. Reference [7] proposes a *Text Cube* model. It is a star schema enriched by a textual dimension defined by a hierarchy; it represents the semantic relationships between the terms of documents. The authors of [8] propose a contextual dimensions modeling by contextual text cube model (*CXT-Cube*). The authors propose two types of contextual dimensions. The *Semantic Dimension* constitutes data extracted from an external knowledge source (generally, domain ontology). The

Metadata dimension represents external metadata of documents (Title, Date, Author, etc.). However, these works focus essentially on the content of documents and ignore their structures.

Others works propose new models. Reference [9] proposes a *Galaxy* model for analyzing XML documents. It is composed by a set of dimensions defined by the user connected by a note instead of Fact. The authors of [2] propose a *Diamond* model that extends the *Galaxy* model by a central dimension (Semantic Dimension); it is connected to other dimensions and represents the semantics of the document. Reference [5] proposes a *CobWeb* model based on a set of facets transformed into dimensions. The authors propose four extensions: exclusion constraints, recursive parameters, duplicated dimensions and correlated dimensions. The main drawback of these works is that they treat the collection of documents having the same structure.

For the OLAP of documents, several works use the Information Retrieval techniques in order to aggregate the textual data.

Reference [11] defines two aggregation functions. The *AVG_KW* function replaces a set of pseudo-average keywords with a smaller and more general set. The *TOP_KW* function determines the k main keywords from the keywords of documents. The authors of [7] propose two measures of aggregation based on Information Retrieval techniques: *terms frequency* (TF) and *inverted index* (IV). Reference [12] proposes two probabilistic measures in order to identify the dominant subject in all documents: *P(wi|topic)* constitutes the word distribution of a topic and *P(topic|dj)* defines the coverage by documents. Finally, Reference [1] proposes essentially three aggregation operators: *List_Concept* (list of the most used concepts), *G_Concept* (the most used generic concepts) and *S_Concept* (the most used specific concepts). The most related works on OLAP of documents propose aggregate functions, measures or operators applied on textual content. This aggregation eliminates infrequent words that could be important for decision makers. In this paper, we integrate the Tag Cloud concept in multidimensional tables in order to highlight the most frequent concepts to decision makers.

3. MULTIDIMENSIONAL ANALYSIS: MODELING

Multidimensional modeling represents an analyzing subject (Fact) according to several axes of analysis (Dimensions). The Star Schema constitutes the fact surrounded by dimensions. Table1 presents the Star Schema Modeling.

Table 1: Star Schema Modeling

Concept	Definition
Star C $= (F ; D)$	F is a non-empty set of $n = 1$ fact. $D = \{D_1,\ldots, D_m\}$ is a set of $m \geq 1$ dimensions.

The fact represents the analyzed subject. Specifically, it models a set of events within an organization (Example: Sales). It is composed of indicators (called measures), generally numeric (see Table II).

Table 2: Fact Modeling

Concept	Definition
Fact F$=(NameF; M_i)$	- *NameF* is the name identifying the fact F - $M = \{M_1,\ldots, M_k\}$ is a set of k measures of F
Measure $\boldsymbol{m_i} = (Name_i; t_i)$	- $Name_i$ is the name of the measure - t_i is the type of the measure

The Dimensions defines the analysis axes of the fact. A dimension is composed of attributes (Parameters and weak attributes). Parameters are organized into hierarchies defined by several levels (different granularities). Hierarchies organize the parameters starting from the finest granularity to finish at the most general granularity. The weak attribute constitutes a descriptive attribute for explaining the semantics of the parameter. Table 3 describes the Dimension Modeling.

Table 3: Dimension Modeling

Concept	Definition
Dimension $\boldsymbol{D_i} = (NameD_i; A_i; H_i)$	- $NameD_i$ is the name identifying the dimension, - $A_i = \{A_{i1},\ldots, A_{iz}\}$ is the set of z dimension attributes (parameters and weak attributes) - $H_i = \{h_1,\ldots, h_{ip}\}$ is the set of p hierarchies showing the arrangement of the attributes of D.
Attribute $\boldsymbol{A_{ij}} = (Name_{ij}; DOM_{ij})$	- $Name_{ij}$ is the name of the attribute, - DOM_{ij} is the domain of the attribute (String, Number…).
Hierarchy $H_{ij} = (P_{ij})$	- $P_{ij} = \{p_{i1} \rightarrow\ldots\rightarrow p_{iy}\}$ is the set of ordered parameters of the hierarchy.

For the multidimensional modeling of XML documents (By content and heterogeneous structure), we propose in this paper a Star Schema composed by a Fact (Keywords) and a set of 5 dimensions (Document, Content, Structure, Semantic and Metadata).

The fact Keywords represents a set of terms extracted from XML documents. We use the techniques of Information Retrieval for determining these keywords.

The dimensions we propose are:

- Document: This dimension links the different information from the other dimensions together.

- Content: The information extracted from the content of documents represents this dimension. We remove the comments, structure, etc.

- Structure: In order to focus on different parts of XML document, this dimension represents the hierarchal structure (it can be determined by DTD, or XML Schema).

- Semantic: This dimension provides to the user the semantics of the content of the XML document. We use the work of [3] for determining this semantics.

- Metadata: This dimension describes XML document by a set of additional data (Example: format, title, rights, etc.).

We present in Table 4 the textual formalism and in Figure 1 the graphical formalism of the Star schema we propose.

Table 4: Multidimensional Modeling of XML Documents: Textual Formalism

Concept	Definition
Star C	Star C = (F, D = {D_1, D_2, D_3, D_4, D_5})
Fact F	F = (Keywords, M = {M_1, M_2}) M_1={Count_KW, Numerical} M_2={List_KW, Textual}
Dimension D_1	D_1=(D_Documents, A={A_{11}, A_{12}}) A_{11}=(IdDoc, Number) A_{12}=(NameDoc, String)
Dimension D_2	D_2=(D_Content, A={A_{21}, A_{22}}) A_{21}=(IdContent, Number) A_{22}=(Content, String)
Dimension D_3	D_3=(D_Structure, A={A_{31}, A_{32}, A_{33}}, H={H_{31}}) A_{31}=(IdElementt, Number) A_{32}=(NameElement, String) A_{33}=(FatherElement, String) H_{31}=(NameElement → FatherElement)
Dimension D_4	D_4=(D_Semantic, A={A_{41}, A_{42}, A_{43}}, H={H_{41}}) A_{41}=(IdConcept, Number) A_{42}=(NameConcept, String) A_{43}=(FatherConcept, String) H_{41}=(NameConcept → FatherConcept)
Dimension D_5	D_5=(D_Metadata, A={A_{51}, A_{52}, A_{53}}) A_{51}=(IdMeta, Number) A_{52}=(NameMeta, String) A_{53}=(ValueMeta, String)

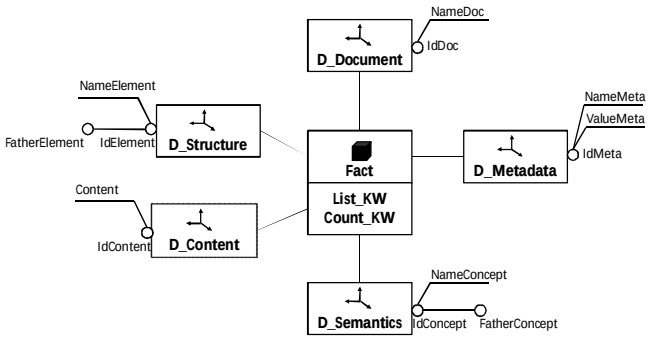

Figure 1: Multidimensional Modeling of XML Documents: Graphical Formalism

3. MULTIDIMENSIONAL ANALYSIS: IMPLEMENTATION

For the OLAP of documents, the content of the multidimensional tables can be numerical (for example, the number of documents, cf. Figure 2) or textual (for example, list of keywords, cf. Figure 3).

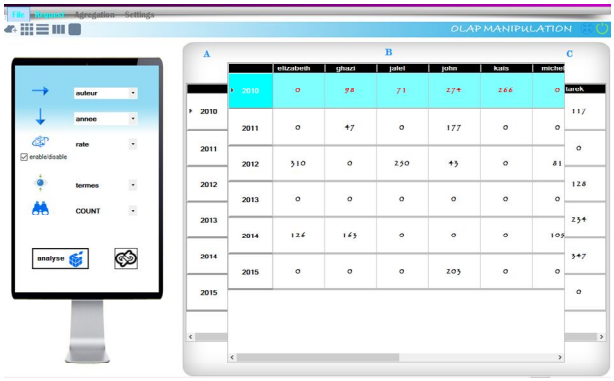

Figure 2: Number of documents by Year and Author

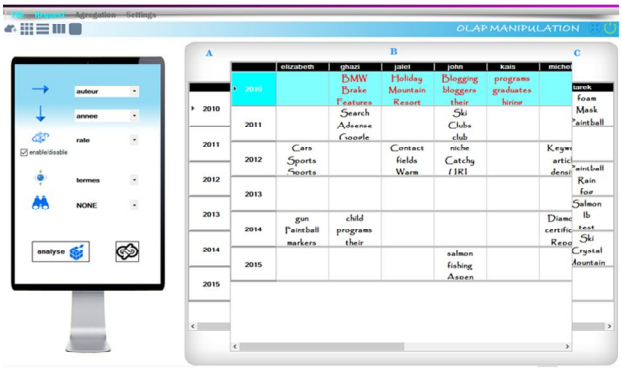

Figure 3: Analysis of Keywords by Year and Author

We note that the content of these multidimensional tables are very congested (too much information). The task of decision makers becomes very complicates. Some works in the literature have proposed aggregation functions, Such *Top_Concept* [1] and *Top_Kewyord* [10]. In this paper, we propose to integrate the concept of Tag Cloud in the multidimensional tables (cf. Figure 4) in order to help users for searching the content easily. In Tag Cloud, the most popular topics are highlighted by using a large font.

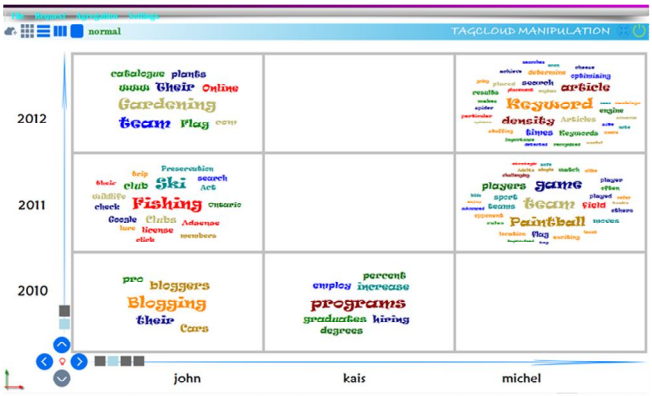

Figure 4: Analysis of Keywords by Year and Author by Using Tag Clouds

In order to draw the attention of decision-makers on the most frequent data by dimensions, we propose two aggregation operators that aggregate the tag clouds by lines or columns (Aggregation by lines, cf. Figure 5 or Aggregation by columns, cf. Figure 6).

For the aggregation function, we calculate the number of occurrences of every word in each dimension axis (Lines or Columns). For every dimension, the relevance of each term is proportional to its frequency in the list of terms of the corresponding dimension.

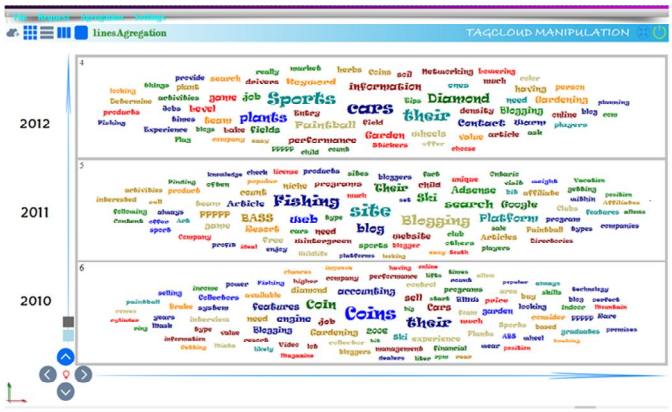

Figure 5: Aggregation by Lines

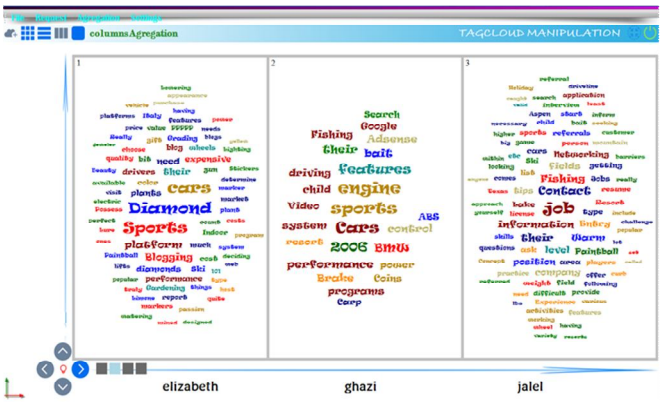

Figure 6: Aggregation by Columns

We propose also another aggregation operator for the total aggregation of documents, i.e. one Tag Cloud (cf. Figure 7).

Figure 7: Total Aggregation

5. CONCLUSION

In order to help the decision-makers to deduce knowledge from the documentary information, we propose a new multidimensional model for XML documents. We have also integrated the concept of Tag Cloud as an appropriate technique for visualizing multidimensional tables. It represents the result of multidimensional query as a set of terms where every term has a font size proportional to its frequency of occurrence.

Several perspectives to this work are possible. We plan to integrate the personalization of analysis OLAP in order to take into consideration the needs of users, based on their profiles. In addition, we propose to introduce the collaborative aspect in order to share of OLAP analyses between different users. Finally, we intend to apply the OLAP techniques on streaming data [6] and big data [4].

REFERENCES

1. M. Azabou, K. Khrouf, J. Feki, N. Vallès, and C. Soulé-Dupuy, **Diamond multidimensional model and aggregation operators for document OLAP**, in *Proc. Research Challenges in Information Science*, Athens, 2015, pp. 363-373. https://doi.org/10.1109/RCIS.2015.7128897
2. M. Azabou, K. Khrouf, J. Feki, C. Soulé-Dupuy, N. Vallès, **Yet Another Multidimensional Model for XML Documents**, *International Journal of Strategic Information Technology and Applications*, Vol. 8, no. 3, pp. 73-90, 2017. https://doi.org/10.4018/IJSITA.2017070105
3. S. Ben Meftah, K. Khrouf, J. Feki, C. Soulé-Dupuy, **A semantic approach for XML document warehousing and OLAP analysis**, *International Journal of Information and Decision Sciences*, Vol. 8, no. 3, pp. 254-283, 2016. https://doi.org/10.1504/IJIDS.2016.078587
4. A. Erraissi, A. Belangour, **Meta-modeling of Big Data management layer**, *International Journal of Emerging Trends in Engineering Research*, Vol. 7, no. 7, pp. 36-43, 2019. https://doi.org/10.30534/ijeter/2019/01772019
5. O. Khrouf, K. Khrouf, J. Feki, **CobWeb Multidimensional Model and Tag-Cloud Operators for OLAP of Documents**, *International Journal of Green Computing*, Vol. 2, pp. 46-68, 2018. https://doi.org/10.4018/IJGC.2018070104
6. K. Kim, J. Song, M. Lee, **Real-time Streaming Data Analysis using Spark**, *International Journal of Emerging Trends in Engineering Research*, Vol. 6, no. 1, pp. 1-5, 2018. https://doi.org/10.30534/ijeter/2018/01612018
7. C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao, **Text cube: Computing in measures for multidimensional text database analysis**, in *Proc. IEEE International Conference on Data Mining*, Pisa, 2008, pp 905-910.
8. L. Oukid, O. Asfari, F. Bentayeb, N. Benblidia and O. Boussaid, **CXT-cube: contextual text cube model and aggregation operator for text OLAP**, in *Proc. Data Warehousing and OLAP*, San Francisco, 2013, pp. 27-32. https://doi.org/10.1145/2513190.2513201
9. F. Ravat, O. Teste, R. Tournier, and G. Zurfluh, **Designing and Implementing OLAP Systems from XML Documents**, *Annals of Information Systems*, Vol. 3, pp. 1-21, 2008. https://doi.org/10.1007/978-0-387-87431-9_15
10. F. Ravat, O. Teste, R. Tournier, and G. Zurfluh, **Finding an Application-Appropriate Model for XML Data Warehouses**, *Information Systems*, Elsevier, Vol. 36, pp. 662-687, 2010. https://doi.org/10.1016/j.is.2009.12.002
11. F. S. C. Tseng and W.-P. Lin., **D-Tree: a multidimensional indexing structure for constructing document warehouses**, *Journal of Information Science and Engineering*, Vol. 22, pp. 819-841, 2006.
12. D. Zhang, C. Zhai, and J. Han, **Topic cube: Topic modeling for OLAP on multidimensional text databases**, in *Proc. SIAM International Conference on Data Mining*, Sparks, 2009, pp 1124-1135. https://doi.org/10.1137/1.9781611972795.96