Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 9, September 2019, pg.24 – 29

Multidimensional Modeling of Semi-Structured Data: XML Documents and Tweets

Kais Khrouf[1,2]

[1]College of Computer and Information Sciences, Jouf University, Saudi Arabia
[2]MIR@CL, University of Sfax, Tunisia

kmkhrouf@ju.edu.sa

Abstract*— The considerable development experienced by the technologies in recent decades has led to the emergence of relatively simple panoply of Internet applications based on open source software, and services designed to improve online collaboration to the large public as: social networking sites (Tweeter, for example) and XML (Extensible Markup Language). Therefore, it is essential to provide efficient tools for decision makers in order to help them analyzing the semi-structured data in a simple way; i.e., as they actually analyze factual or descriptive data. In this paper, we propose a new generic multidimensional model dedicated to the semi-structured data (XML documents and Tweets).

***Keywords*— Multidimensional Modeling, Star Model, Semi-Structured Data, XML Documents, Tweets.**

I. INTRODUCTION

The recent developments of computers and computer networks which connect several systems have facilitated the sharing, communication and access of information for a large number of people. Basically the amount of information produced and exchanged is permanently growing, so users cannot access and handle this information easily; also the information retrieval process becomes painful. Face to this situation, 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 this context, several studies have been interested in the exploitation of semi-structured data (XML documents or tweets).

XML (Extensible Markup Language) has become the standard format for documents. They can be classified into two types: data-centric XML documents and document-centric XML documents. Data-centric XML documents contain short and precise data; in fact, they are very similar to identifiers in relational databases. This type of document is mostly used by applications exchanging information (i.e., transactional data). In such documents, tags precisely describe the content, and then provide the necessary semantic for the comprehension of information contained within the document (e.g., Product, Customer, Quantity, Price are meaningful tags). Whereas, document-centric XML documents are text-rich documents; they constitute the electronic version of traditional paper documents (e.g., scientific articles, internal reports, e-books). Tags used for such documents (e.g., Content, Section, Paragraph) typically describe their logical structure but not their semantics. This is a major drawback in decision support system relying on multidimensional analyses where semantic information is necessary.

Twitter is a system for information sharing; it allows a user to talk about their daily activities, to seek or share information and to track other users who post messages called *Tweets*. A tweet was originally restricted to 140 characters, but on 2017, the limit was doubled to 280 characters for all languages except Japanese, Korean and Chinese. However, the generated code for a tweet is a dozen-line length. In fact, a tweet is a data structure containing several meta-data that could be useful for decision analyses. This structure is composed of mandatory fields visible to twitter users such as the author of the tweet or the tweet's creation date. In addition, other hidden fields dedicated to certain features allow knowing whether the tweet is truncated, or used by the SMS services, its place of issue, or the number of followers, the tweet's unique identifier, etc. Hence, a tweet is not just a text but it can be assimilated to a complex structure including coded information and a collection of meta-data.

The purpose of our work is to propose a new multidimensional model of semi-structured data (XML documents and Tweets), based on complementary dimensions. Each fact includes a set of data and is considered as a means of expression for the users' needs. This paper is organized as follows. Section 2 presents the related work dealing with the multidimensional modeling of XML documents and Tweets. Section 3, describes the basic concepts of multidimensional modeling. Then, we present the new multidimensional model we propose. Finally, we provide the conclusion in Section 5.

II. LITERATURE REVIEW

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

A. XML Documents

The authors in [12] elect the star model for modeling the documents by three types of dimensions: Ordinary (it contains keywords extracted from the document), Metadata (describing the document, e.g., author, language), Category (external data for the document description defined by users according to their viewpoints). Some approaches suggest adding a new specific dimension to the multidimensional modeling of documents. In [14], the authors propose a new model called Topic Cube, based on the star schema; it extends the traditional data cube by integrating a new dimension topics built from the data reflecting ontology of domain and adapted to the preferences of the analyst. The authors of [7] present a cube of text called a Text cube based on the star schema where the textual dimension is represented by a hierarchy of terms. This hierarchy specifies the levels and the semantic relationships between the textual terms extracted of documents.

Other studies have used the metadata and the concepts of domain ontologies as dimensions in order to build the text cubes. For example, (Oukid et al., 2013) [10] propose a contextual text cube model denoted CXT-Cube associated with contextual dimensions which can be classified into two types: i) Semantic dimension which data are extracted from a domain ontology, as an external knowledge source; and ii) Metadata dimension for modeling external metadata of documents such as the date, title and author.

Other work suggest specific models, such as: the Galaxy multidimensional model adapted to the analysis of XML documents [11]. The Galaxy model is based on a unique concept: the dimension concept. This model connects several dimensions by the concept of Node instead of fact. A node associates compatible dimensions for analysis. However, the main drawback of this work is that the authors do not define rules to assist the design phase of a galaxy model. To alleviate this lack, (Azabou et al., 2018) [1] propose a Diamond model, which is the galaxy 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 the proposed model assumes that the collection of documents have the same structure.

(Khrouf et al., 2017) [5] propose a multidimensional model called "CobWeb model" based on standard facets and seen as an extension of the galaxy model [11]. Each facet includes a set of data they transform every facet into a dimension. They propose a set of extensions: i) the exclusion constraint between two dimensions (it requires that a couple of dimensions cannot be used simultaneously in the same analysis), ii) the possibility to define recursive parameters (because the XML documents are represented in a hierarchical manner), iii) duplicated dimension (it is a dimension used more than once in the same analysis) and iv) correlated dimensions (it which allows a same query to move between dimensions).

B. Tweets

A navigation system by facet called NIF-T "Navigating Information Facets on Twitter" based on three facets was proposed in [6]: The Geo Facet allows showing the location of tweets in a map. Subject facet is represented by a word cloud showing the different thematic exchanged by the tweets. This word cloud is generated from tweets by ranking the words on their Term Frequency (TF) scores. Time facet presents the number of tweets in a given date.

(Liu et al., 2013) [8] present a text cube to analyze and model Human, Social and Cultural Behavior (HSCB) from the Twitter stream in a textual database. They are mainly focused on sentiment analysis and visualization. They use methods from behavioral sciences to represent emotion on a numerical measures used in the cube. Positive and negative emotion, anger, anxiety, sadness, religion and social are various measures that the authors use. The authors add two new capabilities to the text cube. The first capability is the heat map functionality that affords a geographical cover of the relevant HSCB derived from linguistic data extracted from the Twitter stream. The second capability is the use of data mining functionality. The text cube architecture supports the development of prediction models.

(Mansmann et al., 2014) [9] propose a system for warehousing Streams from Twitter. Their system lies on an architecture consisting of five layers: i) The data source layer is represented by the available Twitter APIs, ii) The ETL layer (Extract, Transform and Load) for the extraction of data from tweets and processing in a suitable format for the target database, iii) The Data warehouse layer for the storage of data issued from tweets, iv) The Analysis layer dedicated for multidimensional analyses of the tweets, and v) The Presentation layer of analysis results. Nevertheless, the proposed model is inflexible and no scalable.

(Cuzzocrea et al., 2016) [4] define a multidimensional model for the storage of data extracted from tweets streams in order to allow multidimensional analysis. The authors exploit the implicit information issued from the tweets and try to discover the explicitly available metadata. In addition, they propose an aggregation operator for the text integrated in the content of tweets. This operator is based on the *Formal Concept Analysis* (FCA) theory. In second time, they propose a summarization algorithm through multidimensional tweet streams using the concepts of the timed fuzzy lattice structure resulting from *Fuzzy Formal Concept Analysis* [13].

In order to model the tweets' data and to develop an independent application that promotes multidimensional storage of contents of tweets, (Ben Kraiem et al., 2017) [3] propose a multidimensional model dedicated to the content, metadata and social aspect of tweets that is generic and taking into account the structural specificity and possibly semantic data. They retain the constellation schema and then suggest some extensions in order to reflect the specificities of such data. They also extend algebraic operators in order to support analyses by presenting an algebraic formalization and a logical definition as a pseudo code algorithm.

In conclusion, the majority of works dealing with the multidimensional modeling of semi-structured data is concerned by the XML documents or tweets and not both at once. To enhance the modeling capability and analytics of existing models, we propose a new generic multidimensional modeling of XML documents and tweets.

III. MULTIDIMENSIONAL MODELING: BASIC CONCEPTS

Multidimensional modeling consists in considering an analyzed subject as a point in a multidimensional space. The analysts' viewpoint corresponds to a structuring of the data according to several axes of analysis (or dimensions) which can represent various notions such as Time, Geographical location, Products, Suppliers, etc.

A. Analyzed subject

The analyzed subject (i.e., organization's activity as Sales) is represented by the concept of fact. Typically, it models a set of similar events recorded within an organization (e.g., Sales). A fact is composed of a set of measures (indicators) corresponding to the information describing the subject of analysis. These measures are traditionally numeric.

B. Dimensions

The Dimensions represent the axes of multidimensional analysis. They reflect information according to which subjects of analysis will be analyzed. A dimension is composed of attributes. Each attribute represents one data granularity according to which measures could be aggregated. Dimension attributes (also called parameters) are organized into one or more hierarchies, in order to build levels of analysis. A hierarchy is composed of several levels and represents different granularities or degrees of accuracy of information. Hierarchies organize the attributes of a dimension, from the finest granularity to the most general granularity. Thus, a hierarchy defines the valid navigation paths on an analysis axis. The attributes of a dimension that participate in the definition of a hierarchy are called parameters of this dimension. For each parameter, descriptive attributes can be associated. A descriptive attribute is called weak attribute; it is used to explain the semantics of the parameter.

C. Star Schema

A fact and its associated dimensions constitute a Star Schema. It consists of one fact table linked to its associated dimension tables via primary key / foreign key relationships.

Figure 1 depicts an example of a multidimensional model designed as a star schema for the analysis of Sales according to the three dimensions Product, Stores and Time. The fact SALES is composed of two measures Quantity and Amount of sales. The dimension Stores is characterized by three parameters City, Country and

Continent organized hierarchically: The City parameter represents a finer granularity/level than the Country, itself being a finer granularity of Continent.

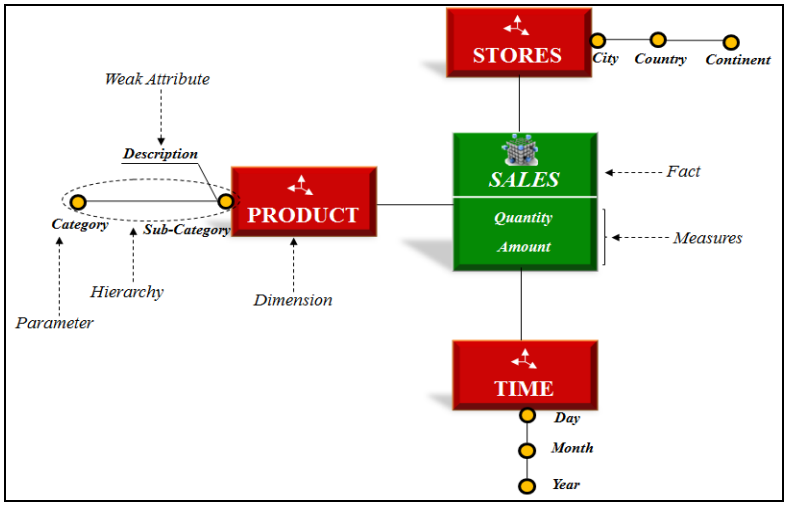

Fig. 1 Example of a multidimensional model: Star schema

IV. MULTIDIMENSIONAL MODELING OF SEMI-STRUCTURED DATA

A. Formalization

We propose the following *Star schema* for multidimensional modeling of semi-structured data (XML Documents and Tweets) in order to reflect the specificities of such data.

Definition 1:

A **star** C is defined by $(F ; D)$ where:
- F is a non-empty set of $n = 1$ fact.
- $D = \{D_1, ..., D_m\}$ is a set of $m \geq 1$ dimensions.

- Fact and its components

Definition 2:

A ***fact*** F is defined by $(NameF ; M_i ; INS_i ; R_i)$ where:
- $NameF$ is the name of the fact F,
- $M_i = \{m_{i1}, ..., m_{ik}\}$ is a set of k measures of F,
- $INS_i = \{ins_{i1}, ..., ins_{il}\}$ is the set of l instances of fact F.

Definition 3:

$\forall$ i $\in$ [1...k], a ***measure*** m_i is defined by $(Name_i; t_i ; f_i)$ where:
- $Name_i$ is the name of the measure,
- t_i is the type of the measure,
- f_i is a set of aggregation functions (SUM, AVG, MAX…).

In order to take into account the specificities of textual data extracted from XML documents or tweets, we distinguish two types of measures: *numerical* measures (has atomic numerical values) and *textual* measures (is a string: one or several words). Table 2 shows the possible aggregate functions by type of measure.

TABLE I
MEASURE TYPES AND THEIR AGGREGATE FUNCTIONS

Type of measure	Aggregate functions allowed
Numerical	Count_KW(Number of Keywords)
Textual	List_KW (List of Keywords)

- Dimensions and its components

Definition 4:

∀ i ∈ [1...m], a ***dimension*** D_i is defined by ($NameD_i$; A_i; H_i) where:
- $NameD_i$ is the name identifying the dimension,
- $A_i = \{a_{i1}, ..., a_{iz}\}$ is the set of *z* dimension attributes (parameters and weak attributes) extracted from *raw data*,
- $H_i = \{h_1, ..., h_{ip}\}$ is the set of *p* hierarchies showing the arrangement of the attributes of *D*.

Definition 5:

∀ *j* ∈ *[1...z], an* ***attribute*** a_{ij} is defined by ($Name_{ij}$; DOM_{ij}) where:
- $Name_{ij}$ is the name of the attribute,
- DOM_{ij} is the domain of the attribute (String, Number…).

Definition 6:

∀ j ∈ [1...p], A ***hierarchy*** h_{ij} is defined by ($Nameh_{ij}$; P_{ij}; $WEAK_{ij}$) where:
- $Nameh_{ij}$ is the name which identifies the hierarchy,
- $P_{ij} = \{p_{i1}, ..., p_{iy}\}$ is the set of parameters of the hierarchy,
- $WEAK_{ij}$: $P_{ij} \rightarrow 2^W$ associates each parameter to a possible empty subset of weak attributes of the dimension of h_{ij}.

B. Graphical Formalism

Based on the concepts defined above, we suggest for the multidimensional model the graphical formalism shown in Figure 15.

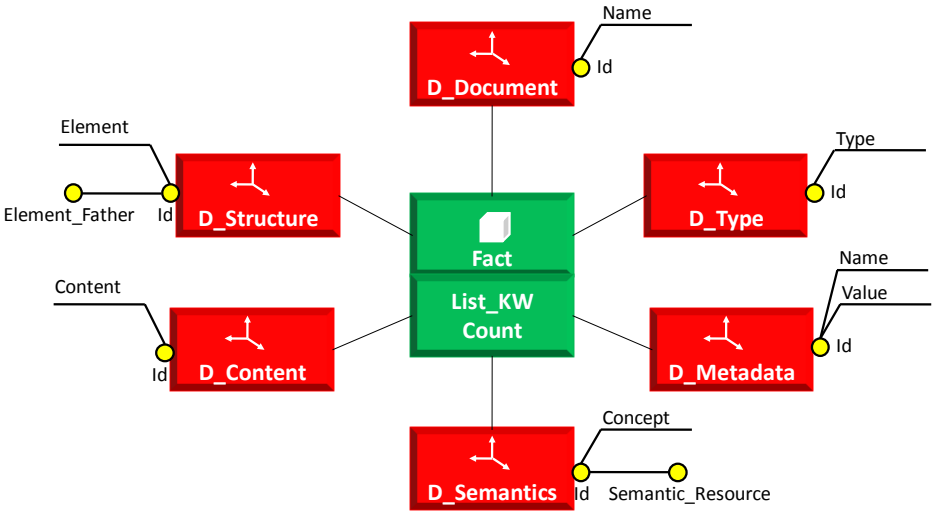

Fig. 2 New multidimensional model for semi-structured data

- Keywords: The fact constitutes a set of the most important Keywords describing the content of the XML document or tweet. These keywords can be determined, by using the indexing techniques of information retrieval, or they come from the document itself when they exist explicitly.
- Structure: this dimension describes the hierarchal structure, it aims to focus on different parts of the XML document or tweet.
- Content: this dimension presents the information contained in the XML document or tweet by removing everything about the comments, structure, etc.
- Semantic: this dimension describes the semantics of the content of the XML document or tweet. 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.
- Metadata: this dimension is provides to the users a set of data describing the XML document or tweet (such as: title, rights, format, dates, etc.).

- Type: This dimension distinguishes the XML documents from tweets. We can analyze the XML documents, the tweets or both at the same time.
- Document: This dimension links the different information from the other dimensions together.

V. CONCLUSION

The multidimensional techniques are broadly used in various application domains. Their capability to handle aggregation analysis makes it worthwhile to investigate those technologies for analyzing semi-structured data. However, these data lays different challenges, i.e., format, volume and speed data that should be considered to enable multidimensional analysis. In this paper, we propose a new generic multidimensional model of semi-structured data (XML documents and Tweets). Several perspectives to this work are possible. At first, it is important to propose new multidimensional operators that take into consideration the specificities of the proposed multidimensional model. We intend thereafter to integrate non-textual contents, e.g., photos, videos, audios, etc., from semi-structured data into the multidimensional perspective.

REFERENCES

[1] 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*, Volume 8, Issue 3, pp. 73-90, 2017.

[2] 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*, Volume 8, Issue 3, pp. 254-283, 2016.

[3] M. Ben Kraiem, J. Feki, K. Khrouf, F. Ravat and O. Teste, "OLAP operators for missing data". *Business Intelligence & Big Data*, Volume B-13, pp. 53-66, 2017.

[4] A. Cuzzocrea, C. Maio, G. Fenza, L. Vincenzo and P. Mimmo, "OLAP analysis of multidimensional tweet streams for supporting advanced analytics", in *31st Annual ACM Symposium on Applied Computing*, 2016, pp. 992–999.

[5] O. Khrouf, K. Khrouf, J. Feki, "CobWeb Multidimensional Model and Tag-Cloud Operators for OLAP of Documents", *International Journal of Green Computing*, Volume 2, pp. 46-68, 2018.

[6] S. Kumar, F. Morstatter, G. Marshall, H. Liu, and U. Nambiar, "Navigating Information Facets on Twitter (NIF-T) ", in Proc *ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1548-1551.

[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*, 2008, pp 905-910.

[8] X. Liu , K. Tang, J. Hancock, J. Han, M. Song, R. Xu and B. Pokorny, "A text cube approach to human, social and cultural behavior in the twitter stream", *Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 321-330, 2013.

[9] S. Mansmann, N. Rehman, A. Weiler, M.H. Scholl, "Discovering OLAP dimensions in semi-structured data", *Information Systems*, Volume 44, pp. 120-133, 2014.

[10] 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*, 2013, pp. 27-32.

[11] F. Ravat, O. Teste, R. Tournier, and G. Zurfluh, "Designing and Implementing OLAP Systems from XML Documents", *Annals of Information Systems*, Volume 3, pp. 1-21, 2008.

[12] F. S. C. Tseng and W.-P. Lin., "D-Tree: a multidimensional indexing structure for constructing document warehouses", *Journal of Information Science and Engineering*, Volume 22, pp. 819-841, 2006.

[13] R. Wille, "Restructuring lattice theory: An approach based on hierarchies of concepts". In Proc. *International Conference on Formal Concept Analysis*, 2009, pp. 314.

[14] 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*, 2009, pp 1124-1135.

