Generic Multidimensional Model of Complex Data: Design and Implementation

Kais Khrouf and Hela Turki
Kmkhrouf@ju.edu.sa, Hnalturki@ju.edu.sa
Jouf University, King Khaled Road, Al Jouf, Saudi Arabia

Summary
The use of data analysis on large volumes of data constitutes a challenge for deducting knowledge and new information. Data can be heterogeneous and complex: Semi-structured data (Example: XML), Data from social networks (Example: Tweets) and Factual data (Example: Spreading of Covid-19). In this paper, we propose a generic multidimensional model in order to analyze complex data, according to several dimensions.

Key words:
Multidimensional Analysis, Star Schema, Complex Data, XML, Tweets, Coronavirus Covid-19.

1. Introduction

The data analysis techniques provide statistics and useful information (presented in graphs, charts, tables...) in order to reduce the risks of decision-making; it is the process of retrieving, cleaning and loading data from heterogeneous sources, and extracting relevant information and knowledge that helps decision-makers to make the right decisions. So, data analysis offers better ways to analyze and study business and scientific data.

In the literature, there are many data analysis techniques. We propose in this paper to use the Multidimensional Analysis techniques to study complex data (semi-structured data as XML, data from social networks as Tweets and factual data as Coronavirus Covid-19).

Multidimensional analysis is a technique that allows decision-makers to assess data from different points of view; the information then is organized by hierarchy, in order to facilitate the analysis of data. A hierarchy is a set of ordered levels (analysis unit) of a dimension (axis of analysis).

In this paper, we propose a Generic Multidimensional Model based on Star Schema to analyze complex data. We present models generated from the proposed generic multidimensional model for: XML data, Tweets data and Coronavirus Covid-19 data.

This paper is organized as follows. In Section 2, we present the literature review of works on multidimensional analysis techniques. The Section 3 describes the generic multidimensional model we propose for complex data. In next Sections, we present instantiations of proposed generic model: the first example concerns XML data, the second example focuses on data from Tweets and the last example treats data of Coronavirus Covid-19 spreading.

2. Literature Review

Multidimensional Analysis techniques or On-Line Analytical Processing (OLAP) tools have been proposed in order to help decision-makers by exploring useful data according to several perspectives.

The authors of [2] propose a multidimensional model for analyzing data of health services to enhance information extracted from several data sources. They propose to discover trends in health resources and medical records. The authors of [1] propose a data warehouse to examine the correlation between (1) the Coronavirus covid-19 spreading data and (2) pollution and climate data. This study shows that regions characterized by the absence of rain and wind are significantly contaminated by the novel virus.

In [3], the authors propose a diamond model by adding a central dimension; it is connected to other dimensions and it describes the semantics of documents (list of concepts). In [6], the authors propose to compress XML data into OLAP cubes. They present a multidimensional model based on snowflake schema; it is composed by several XML documents (every document relates to a dimension) and a single XML fact document.

By taking into account the social aspect of tweets, the authors of [7] calculate the mention frequency by users and estimate for each event the time period in order to detect events on Twitter. The authors of [4] use a constellation schema (several facts surrounded by common dimensions) in order to design a multidimensional model of data extracted from tweets. They analyze the content

Manuscript received December 5, 2021
Manuscript revised December 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.12.88

and the metadata of Tweets and they propose a new concept: Reflexive Fact.

In this paper, we propose a Generic Multidimensional Model based on Star Schema to analyze complex data: XML data, Tweets data and Coronavirus Covid-19 data.

3. Generic Multidimensional Model

A Data Warehouse (DW) is a repository used for analyzing and reporting data integrated from multiple disparate sources; it stores historical data contrary to the company's operational systems. A Data warehouse provides consolidated data through the interactive OLAP (Online Analytical Processing) tools [5].

Multidimensional modeling presents data in the form of data cubes (or Multidimensional Tables); it models and views the data (central theme) according several perspectives (dimensions). In the literature, several schemas were proposed. We focus on the simplest and the most used schema, namely Star Schema.

A star schema constitutes a data warehousing model where one fact is surrounded by multiple dimension tables; it is called star schema because the diagram of this schema resembles a star: Dimension Tables radiating from the central table: Fact Table. A star schema is represented by a large Fact Table that describes metrics in business process and a set of smaller Dimension Tables (Descriptive information).

Star Schema SC = $(FT ; DT_i)$
FT is the Table Fact.
DT_i is a set of Dimension Tables.

A *Fact Table* is considered as central table in a star schema; it is denormalized and stores quantitative information (measures for a specific event) for analysis. A fact table is described by measures (set of indicators).

Fact Table FT=$(NameFT; MS_i)$
$NameFT$ is the name of the Fact Table FT.
MS_i is a set of measures describing the Fact Table.

A *Dimension Table* contains attributes (and weak attributes) that describe the objects in a fact table; it is a collection of references about the event (Fact Table). A Dimension table constitutes the analysis axis; it represents descriptive information.

Dimension Table DT$_i$ = $(NameDT_i; ATT_j; HIERAR_k)$
$NameDT_i$ is the name of the Dimension Table,
ATT_j is the set of attributes (weak attribute or parameter).
$HIERAR_k$ is the set of hierarchies for the Dimension Table.

4. Multidimensional Model of XML Documents

XML (eXtensible Markup Language) is a semi-structured data that facilities the exchange the data on the Web. We distinguish two types of XML documents: Text-oriented XML documents, such as: reports, papers; and data-oriented XML documents, such as: Documents generated from Excel or databases.

For this type of complex data, we generate the following multidimensional model (Star Schema) composed by a Fact (list of keywords) surrounded by a set of five dimensions (Document, Content, Structure, Semantic and Metadata).

We use the Information Retrieval techniques for extracting the list of keywords from XML documents and PERL Parser to instantiate the proposed model.

The dimensions are: Document (characterized by an id and its name), Content (the extracted information from XML documents), Structure (hierarchal structure, determined by DTD, or XML Schema), Semantics (Knowledge extracted from content) and Metadata (set of additional data).

Fig. 2 presents the proposed multidimensional model of XML documents.

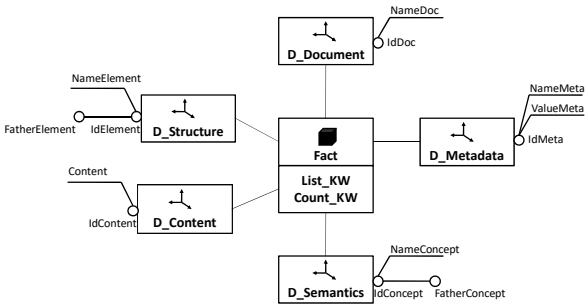

Fig. 1 Multidimensional Model of XML Documents.

The next figure presents the list of keywords by year and author.

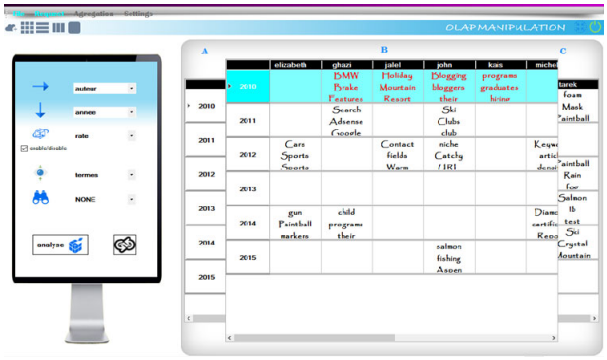

Fig. 2 List of Terms by Dimensions (Year and Author)

This multidimensional table is congested. Some aggregation functions (*Top_Concept* [2], *Top_Kewyord* [8]) were proposed in the literature. In order to help decision makers (interpret easily the content of the multidimensional table), we use the concept of Tag Cloud; it visualizes the terms as big as their frequency is high (cf. Fig. 4).

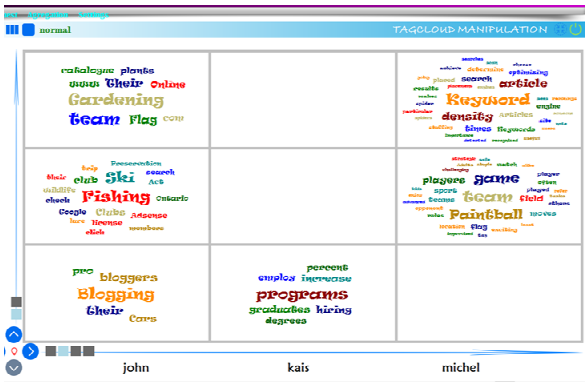

Figure 3 Analysis of Terms by Using Tag Clouds

5. Multidimensional Model of Tweets

A tweet is composed of two parts: the first part (short message less than 140 characters) visible to users and the second part (generated code containing some lines) hidden to users. A tweet constitutes a data structure with information (User and Meta). The visible fields can be the author, the creation date, but the hidden fields can be the tweet's ID, the place of issue, etc.

A data structure of tweets contains three parts:
The Tweet part includes the id, the text of the tweet, the creation date, the number of re-tweet, the source (Such as: Web, Twitter for Android), the ID of an existing tweet (If it is a tweet response), the screen name and the user ID of tweet replied. The User part includes the information about the user: the ID, his name, his screen name, the language and the Time_Zone. The place part presents the name, the type (City or Neighborhood) and the country.

The proposed fact is composed of two measures: one textual (the text of Tweet) and one numeric (Number of retweet). The proposed dimensions are: D_Source, D_Time, D_Place and D_User.

Fig. 4 presents the proposed multidimensional model of Tweets.

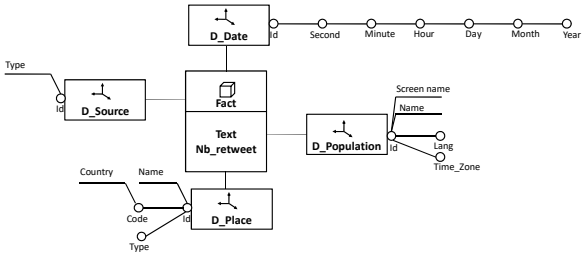

Fig. 4 Multidimensional Model of Tweets.

Table 1 present the Number of tweets per time-zone and source.

Table 1: Number of tweets per time-zone and source

2

Time-Zone	Source			
	Twitter for Android	*Twitter for BlackBerry®*	*Twitter for iPhone*	*Web*
Central Time (US & Canada)	925	332	1766	1214
Brasilia	334	126	178	2638
Eastern Time (US & Canada)	709	204	1710	1323
Santiago	213	133	164	1366
Quito	489	226	976	603
Greenland	197	-	134	837
Pacific Time (US & Canada)	454	556	788	566
Hawaii	340	224	426	447
Amsterdam	170	116	253	245
Atlantic Time (Canada)	377	-	782	500
Baghdad	119	161	-	224
London	150	136	420	310
Mountain Time (US & Canada)	221	138	418	312
Tokyo	206	-	353	134

6. Multidimensional of Coronavirus Covid-19 Data

The epidemics and pandemics, such as Spanish flu in 1917 and H1N1 in 2009 constituted major problems of the humanity. In January 2020, WHO (World Health Organization) has detected a novel Coronavirus, called CoVID-19. In March 2020, it becomes a global pandemic, according to declarations of the WHO [9].

In order to study the evolution and spreading of this new virus, we propose the multidimensional model that contains four dimensions:
- D_Date: includes the following hierarchy: Date, Month and Year. The weak attribute Day is associated to the parameter Date.
- D_Country: described by the following hierarchy: Code, Zone and Continent. We associate the weak attribute Country to parameter Code.
- D_Sex contains the parameter Code and the weak attribute Sex.
- D_Population: contains the parameter Id and the weak attribute Interval.

Fig. 5 presents the proposed multidimensional model of CoVID-19 data. The measures of the fact are: Cases and Deaths. The four dimensions of this multidimensional model are: D_Date, D_Country, D_Population and D_Sex.

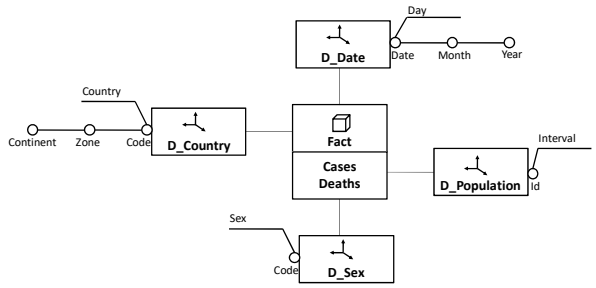

Fig. 5 Multidimensional Model of CoVID-19 Data

Table 2: Average of Positivity Rate by Region and Trimester

	T1-2021	T2-2021	T3-2021
Eastern Europe	18,29	9,43	0,71
Northern Europe	3,34	2,00	1,81
Southern Europe	7,73	3,87	4,19
Western Europe	5,29	3,78	2,79

Table 3: Average of Positivity Rate by Country and Trimester

	T1-2021	T2-2021	T3-2021
Austria	2,22	0,40	0,18
Belgium	5,74	5,65	3,27
Bulgaria	13,01	8,41	3,04
Croatia	10,34	11,83	1,98
Cyprus	1,13	0,72	0,99
Czechia	11,93	0,94	0,22
Denmark	0,52	0,15	0,41
Estonia	14,93	6,98	4,44
Finland	2,66	1,28	2,46
France	6,68	4,91	3,98
Germany	6,96	6,38	3,88
Greece	3,63	0,87	1,54
Hungary	14,78	8,50	1,10
Iceland	0,38	0,26	1,56
Ireland	7,70	2,45	5,88
Italy	6,90	3,54	2,16
Latvia	6,61	3,45	0,93
Lithuania	8,83	4,35	3,55
Luxembuurg	1,92	1,43	1,42
Malta	6,06	1,19	2,94
Netherlands	9,46	8,13	4,07
Norway	1,38	1,44	1,40
Poland	18,82	10,86	0,52
Portugal	6,79	1,11	2,68
Romania	13,53	5,61	1,47
Slovakia	46,43	8,03	0,70
Slovenia	6,06	1,93	0,95
Spain	8,81	4,96	9,34
Sweden	11,76	8,25	3,60

5. Conclusion

In this paper, we propose a generic multidimensional model based on Star Schema (a central fact surrounded by a set of dimensions) in order to deduce knowledge from complex data. We distinguish three types of complex data: Semi-structured data as XML, Data from social network as Tweets and factual data as Covid-19 Spreading.

For future work, we propose to add constraints or criteria on multidimensional tables based on colors in order to highlight the most important values.

References

[1] G. Agapito, C. Zucco, M. Cannataro, “COVID-WAREHOUSE: A Data Warehouse of Italian COVID-19, Pollution, and Climate Data”, *International Journal of Environmental Research and Public Health*, Vol. 17, No 15, 2020.

[2] B. Appah, D. Amos, “Multidimensional Data Model for Health Service Decision Making Data”, *International Journal of Computer Science Engineering Techniques*, Vol. 3, No. 3, P. 1-6, 2018.

[3] M. Azabou, K. Khrouf, J. Feki, N. Vallès, C. Soulé-Dupuy, “Diamond multidimensional model and aggregation operators for document OLAP”, *Research Challenges in Information Science*, Athens, 2015, pp. 363-373.

[4] M. Ben Kraiem, M. Alqarni, J. Feki, F. ravat, "OLAP operators for social network analysis", *Cluster Computing*, 2019.
[5] T. Berber Sardinha, M. Veirano Pinto, Multi-Dimensional Analysis Research Methods and Current Issues, Bloomsbury Academic, 2019
[6] D. Boukraa, M. A. Bouchoukh, O. Boussaid, "Efficient Compression and Storage of XML OLAP Cubes", *International Journal of Data Warehousing and Mining*, Vol. 11, No 3, 2015
[7] A. Guille, C. Favre, "Event Detection, Tracking and Visualization in Twitter: A Mention-anomaly-based Approach", *Social Network Analysis and Mining*, Vol. 5, No. 1, 2015.
[8] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, "Finding an Application-Appropriate Model for XML Data Warehouses", *Information Systems*, Elsevier, Vol. 36, 2010.
[9] https://www.who.int/health-topics/coronavirus

Dr. Kais Khrouf was born in Paris, France, 1977. He received the B.S. degree in Computer Sciences from University of Sfax, Tunisia in 1999. He received the M.S. and Ph.D. degrees in Computer Sciences from University of Paul Sabatier, Toulouse, France. From 2005 to 2017, he was an assistant professor at University of Sfax, Tunisia. Since 2017, he is an assistant professor at Jouf University, Saudi Arabia. He is the author of more than 40 articles in international journals and conferences and he is a permanent reviewer in International Journal of Information and Decision Sciences (Inderscience Publishers). His research interests include Decision Support Systems, Data Analysis, Data Warehouses, Social Networks and Semi-Structured Data.

Dr. Hela Turki received the B.S. and the M.S. degree from University of Sfax, Tunisia in 2012. She received the Ph.D. degree from University of Sfax, Tunisia in 2017. From 2015 to 2017, she was a contractual professor at University of Sfax, Tunisia. Since 2017, she is an assistant professor at Jouf University, Saudi Arabia. She is the author of many papers in international journals and conferences. Her research interest includes Decision Making and Data Analysis.