Revista de Sistemas de Informação da FSMA
n. 14 (2014) pp. 26-35

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL

Considering unstructured data for OLAP: a feasibility study using a systematic review

Sahudy Montenegro González, Talita dos Reis Lopes Berbel
Centro de Ciências em Gestão e Tecnologia (CCGT) -
Universidade Federal de São Carlos (UFSCAR). Sorocaba - Brasil.
sahudy@ufscar.br, talitareislopes@gmail.com

***Abstract*—Among the technologies involved on Business Intelligence, Data Warehouse and OLAP have been widely used to identify, collect, process, integrate and analyze information for decision making, thus promoting business management. Data stored in a data warehouse used by conventional OLAP systems are structured in nature. However, data such as text documents, images and videos, characterized as semi or unstructured data, may also contain information of great value to business. In this context, we applied the systematic review methodology with the purpose of identifying, extracting and summarizing the main research results focused on the integration of unstructured data in data warehouse environments. We raised forty two studies which were classified in order to identify ongoing research subjects and potential gaps as future trends.**

***Index Terms*—Data warehouse, multidimensional modeling, OLAP, systematic review, unstructured data.**

I. Introduction

TRADITIONAL OLAP (OnLine Analytical Processing) was not meant to process large volumes of unstructured data and traditional Data Warehouse (DW) was not designed to store them.

The advent of the Web and other technologies has increased the volume of unstructured data. Data such as text, images and video characterized as semi or unstructured data, may contain information of great value to business and the decision making process. Reports, emails, customer complaints and suggestions, video surveillance, voice on customer calls can be used to generate knowledge. These kinds of data can be easily accessible and are an attractive source of information, but the process of extract and search for information has major challenges because these types of data are schema-less or self-describing and do not have a predefined data model. They are heterogeneous, voluminous, variable in nature and have different formats.

The volume of unstructured data grows faster than the volume of structured. According to a 2011 study by IDC (International Data Corporation), unstructured data will be responsible for 90% of all data created in the next decade.

The ideal process of integrating unstructured data on the context of data warehouses is to manage, query and visualize information in a way that is as effective and meaningful with structured data. The integration of unstructured with structured data can produce a powerful source of information and decision-making.

According to [47], "*when the gap between unstructured data and structured data is bridged, an entirely new world of possibilities and opportunity for information systems opens up*".

Given this motivation, this work describes a systematic review that aims to justify that demand, in order to understand, analyze and synthesize the proposals and, thus to identify current researches and future policies. This paper follows the systematic review guidelines described in [40].

We extracted forty-two scientific articles from journals and conference proceedings and performed a qualitative assessment to classify the studies by subjects in order to identify ongoing research topics and potential gaps as future trends.

In this paper, we study only a specific part of the process of integrating unstructured data into data warehouses, that is, ETL proposals that deal with unstructured data sources but traditional data warehouses are beyond the scope of this review. The purpose is to highlight studies with new models and operations to manage unstructured data in data warehouses for OLAP.

The paper is organized as follows. Section 2 briefly outlines the systematic review process. Section 3 presents the systematic review and the list of selected articles. Section 4 presents a description of the forty-two studies classified by topics: multidimensional modeling and data warehouse design for multimedia and text document data; data extraction, cleansing, transformation and loading; data warehouse architecture; analytical front-end tools and maintenance and evolution of data warehouses. The evaluation process includes a summary to show evidences of implementation, the existence of case studies and/or patents and provenance (academic or industry) for each paper under study. Finally, in Section 5 we discuss the overall results and possible future directions.

I. CONDUCTING THE SYSTEMATIC REVIEW

In order to survey the literature concerning the integration of unstructured data into data warehouses, we carried out a systematic review. There has been increasing interest on this topic listed as one of the next challenges on the field as presented in [45].

A systematic review aims to identify, evaluate and interpret relevant results for a research topic or question. It consists of three phases: planning, execution and synthesis.

The planning of the systematic review defines the protocol including the research goals, research question and the methods to conduct the review.

In the execution phase, it is possible to identify and evaluate studies by using inclusion and exclusion criteria leading to the final selection.

Finally, the summarization process allows us to understand the current state of the proposals towards the research goals and answer the research questions. The final stage of this systematic review took place on February of 2014.

A. Research Question

In order to clearly define the goal of the systematic review, we defined the following research question: "*What initiatives explicitly consider unstructured or semi-structured data for Data Warehouse and OLAP?*" By unstructured data we consider multimedia or document data and by semi-structured we are mentioning text document data.

Keywords were carefully chosen to represent the research question. Terms related to the research question are divided into three categories: problem, mechanism and measurement. In order to execute the systematic review we used the following search expression:

```
PROBLEM
("unstructured data" OR "semi-structured data" OR
"multimedia" OR "document" OR "text")
AND
MECHANISM
("warehouse" OR "warehousing" OR "OLAP" OR
"analytical processing")
AND
MEASUREMENT
("model" OR "technique" OR "platform "OR" framework
"OR" architecture "OR" implementation "OR" system
"OR" application "OR" tool "OR" method "OR" review
"OR" approach "OR" proposal "OR" experience ")
```

As our research is being conducted on a recent topic of investigation, we included not only implemented proposals but also models and previous reviews. The publications were in the field of data warehouse and OLAP. The expected outcome was to identify and classify the initiatives related to Data Warehouse with unstructured data.

The bibliographic databases used to perform the search were Web of Knowledge, Science Direct and Scopus. These databases include the most relevant publishers, such as IEEE, Elsevier, Springer and ACM publications.

Search engines have different interfaces; hence, a specific query string was built for each database based on the expression above.

B. Selection of the Primary Studies

The two authors independently selected studies from among all the search results. In order to complete the protocol, the strategies to conduct the review and eliminate disambiguation and disagreement were established, so that we could finish the first phase of this work, the determination of selection criteria. The criteria for inclusion of articles in the review were:

(1) studies on initiatives that have been undertaken specifically and explicitly on Data Warehouse and OLAP with semi-structured or unstructured data (multimedia and/or text document data);

(2) articles whose keywords are in the search string or who include at least one word from each category of the search string in their title or abstract; and

(3) articles written in English.

The initial criteria for exclusion of articles were:

(1) those whose main focus was not either data warehouse or OLAP;

(2) those whose full text was not available through *CAPES*[1] *Portal de Periódicos*;

(3) those whose author's previous work had already been included in the systematic review. It is common practice to publish a research at various stages. Hence, we verified and selected the newest publication since it was probably the one with the most complete description and results. This criterion is not applied to complementary articles of the same authorship, for example, when one paper presented the cube model and the other focused on the ETL process. The review contemplated the latter case as two different studies from the same authors.

(4) Those reviews that were repeated afterwards with the same goals. In this case, we selected the most recent review about each subject related to the research question.

During the selection of primary studies, we found several articles exploring unstructured data in the ETL process. Hence, we included one more exclusion criterion to complete the protocol: proposals dealing with traditional data warehouses. This refers to studies whose ETL processes manipulated unstructured data sources but solely stored traditional data types without semantics or any representation of their nature.

C. Qualitative Criteria to Evaluate Primary Studies

In order to provide a perspective of the studies in the

[1] http://www.periodicos.capes.gov.br

summarization phase, we decided to put together similar features into categories.

Let us first start with a brief description of the overall process to achieve this goal.

Unstructured-centric collections require specific data cube and visualization models, architecture and front-end tools that can handle structural and semantic constraints. Traditionally, as part of a data warehouse architecture, the ETL process must clean, transform and load/update data on the warehouse.

For unstructured data, the ETL process uses information extraction, content-based information techniques, among others, to perform the processes tasks.

The data stored into the data warehouse can either be subject to standard OLAP or be described by specific data model and include new OLAP operators.

Generally, unstructured data can be represented with multiple interpretations. For example, multiple semantic constraints can be extracted from a document and multiple features can be extracted from image content (color, texture, etc.). The dynamic nature of these data demands the search for resources to ensure maintenance and evolution of the data schemas and representation.

We narrowed the main research subjects on data warehouse and OLAP for semi-structured and unstructured data under the following approaches:

• *Document-centric approach*: studies focused on storing, managing and/or providing support to semi-structured or unstructured data, also called document or textual data. The model is centered on text documents.

• *Multimedia-centric approach*: studies focused on storing, managing and/or providing support to unstructured multimedia data, such as audio, video and images. The model is centered on one or more multimedia data sources.

We then defined five different main topics (**Class A** to **Class E**) to describe document-centric and multimedia-centric approaches:

• **Class A**- *Multidimensional modeling and Data Warehouse design*: studies focusing on conceptual solutions for multidimensional modelling, which include cube modelling and algorithms for OLAP. Logical and physical designs such as index structures were also considered under this topic.

• **Class B**- *Data extraction, cleansing, transformation and loading*: studies focusing on ETL process (extract, transform and load) and whose data integration was performed from multiple and different data sources. The process includes data cleaning, transformation and finally loading the data into a data warehouse or datamart. ETL proposals that dealt with unstructured data sources over traditional data warehouse environments were not considered as explained in the previous subsection (*Selection of Primary Studies*).

• **Class C**- *Data Warehouse architecture*: studies focusing on the DW architecture for the integration of semi-structured and/or unstructured data.

• **Class D**- *Analytical front-end tools*: studies focusing on DW applications that allowed end users and business experts to create their queries involving semi-structured and/or unstructured data.

• **Class E**- *Maintenance and evolution of Data Warehouses*: studies focusing on simplifying maintenance and evolution of DW schemas and therefore, future corrections, development, reuse and management.

It is possible that the studies be included in one or more of the above categories.

Besides, we defined three criteria to emphasize the quality of the studies. Thus, each study was summarized according to whether:

1. *the study presented evidence of implementation*: the study reported a prototype or whether it was merely a theoretical proposal. This category does not demand the existence of a case study;

2. *the study presented a case study*: the study explicitly reported an empirical research with implementation and experimental results. In this case, it should be clear that there was an implementation created beforehand;

3. the study is an academic initiative or an industry application (*provenance*) and whether the experience may result in patents.

II. Executing the Systematic Review

Figure 1 illustrates the study selection process. Our search terms identified a total of 474 publications including 47 duplicated articles from all databases. During Phase 2, the remaining 427 studies were analyzed by title, keywords and abstract to determine whether they were in accordance with the research question, reducing the collection to 72 studies. Phase 3 reviewed the publication contents to determine their relevance to the research question.

At the end of this process, forty two studies were found to be in compliance with all inclusion/exclusion criteria. In the case where the abstract was not enough to understand the goals of the proposal and to determine its adequacy the full text was then examined.

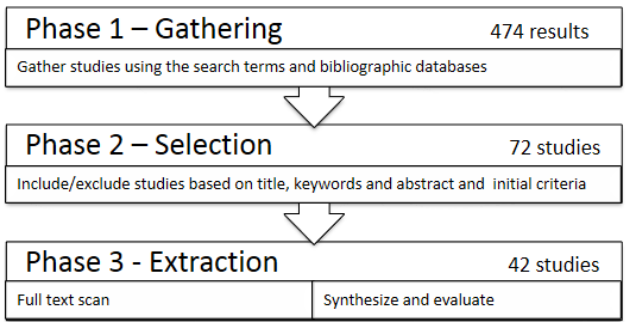

Fig. 1. Stages of the study selection process

The extraction phase of the systematic review accepted 57% of the selected studies on Phase 2. This means that 30 studies were rejected according to the exclusion criteria. The causes of rejections were:

- approach out of the scope of our study (38.7%);

- full text not found (29%);
- previous similar work from the same authors (32.2%).

Some of the reasons to exclude twelve articles were:

(1) the use of the word "warehouse" referring to data repositories; or

(2) the description of document or multimedia models with no mention to data warehouse or OLAP.

The lack of online availability of some publications did not prevent us from covering the major proposals. We consider that the results provided us with an adequate view of the state of the art of this community. Other criteria of exclusion were not applied in this phase.

To support our review we used the StArt (State of the Art through Systematic Reviews) tool [17], developed at the Federal University of São Carlos, Brazil, to assist the process of systematic reviews.

III. QUALITATIVE ASSESSMENT

In the next subsections, we summarize the studies according to the main research subjects specified in the previous sections. We will describe the main contribution of each work in each category. There are some works that will be present in more than one category.

A. Multidimensional Modeling and Data Warehouse Design

In this section we describe the contributions for document-centric and multimedia-centric approaches.

Document-centric:

The term Document Warehouse, a framework for analysis, sharing and reuse of unstructured data, emerged in several studies.

In the context of document warehouses, there are three main techniques used to extract information from text data: text mining, information retrieval and information extraction.

Text mining has become an increasingly popular key, because the traditional techniques of information retrieval became inadequate for large amount of text data [18]. Systems using text mining are able to extract keywords and summarize texts, thus allowing text classification or similarity clustering such as the one described in [26], which describes categorization techniques used to associate text documents to dimensions as part of the DocCube system.

The DocCube model was based on the star schema and introduced concept hierarchies to structure a document collection.

In [20] it was presented a cube model and its OLAP operations for text documents. The Index Cube is a three-dimensional cube using three index structures. The measures were information retrieval and text mining measures, such as Term Frequency (TF) and document frequency.

Text-Cube [22] is a cube model that defined two measures and one special type of dimension for text data. The measures aggregated text data also using term frequency and Inverse Document Frequency (IDF) weights. The term hierarchy in the special dimension specified semantic levels and relationships among text terms.

A data cube model called Topic-Cube was proposed in [50]. It extended the traditional data cube to make it able to process topic hierarchies and defined two types of probabilistic measures based on topics: word distribution of a topic and topic coverage by documents.

iNextCube [49] integrates the Text-Cube [22] and Topic-Cube [50] models with information network analysis into an architecture to query and mine multidimensional text databases. Information network analysis was used to generate dimension hierarchies.

The proposal presented in [16] is keyword-centric for OLAP queries. They proposed two measures, total frequency and total documents, both ranked by keywords.

The authors of [37] presented a model named Galaxy for analyzing XML documents. A galaxy has a set of dimensions and a function that determines compatibility between dimensions. The authors provided a different modeling approach, which consisted in a multidimensional model without fact table. This model was based on dimensions and its aggregations are performed using the Top_Keyword aggregation function presented with details in [36].

Top_Keyword is based on the information retrieval TF-IDF metric. In addition, in [36] a different model was described including textual measures on the fact table and Top_Keyword is applied on these measures.

The study made in [34] used the concept of Galaxy [37] associated to an adapted design process to specify analysis over XML documents. The model created multidimensional structures representing user requirements.

The work described in [8] proposed a logical model for analyzing XML documents using an abstract XML tree model. and described new operators to support structured aggregation over XML data.

The authors of [43] proposed a set of ten definitions about document, dimension, document index, cell and document cube to formalize document warehousing for multidimensional analysis. They projected three types of dimension (ordinary, metadata and category) and a document counter as the default measure in a document cube.

The paper in [44] complemented [43] describing an indexing structure, called D-Tree, which is a tree structure to implement a document cube. It was used as metadata for document warehousing.

In [33]; [31] it was defined the contextualized warehouse, which performed analysis on an OLAP cube, called R-cube. The R-cube multidimensional data model aimed to retrieve documents and facts relevant for the selected context, by maintaning two special dimensions: relevance and context. This approach used information retrieval and probabilistic models and proposed Algebra operations to manage R-cubes.

The study described in [7] proposed a Document Concept Graph (DCG) - a generic data model to store information extraction contents, an OLAP Model and mapping steps for transformation between them. The user entered entity types and entities of interest (posted as queries) to be extracted from unstructured text. The semantic relationships intra and inter documents were stored into the graph model. Then, an OLAP model was populated (instances and metadata) from a DCG to analyze the data by using Business Intelligence tools.

The authors of [9] created five definitions to build their multidimensional model of complex objects previously proposed in [10]. They defined a set of OLAP operators (1) to construct complex data cubes and (2) to visualize data cubes. The former included the cubic projection and cube-based operators. The cubic projection builds a complex cube from a multidimensional schema and cube-based operators create new cubes from existing ones, while the cube-based operator comprised the view projection operation, the aggregate operators and "slice-and-dice like" operators, which manipulate features or measures according to the user query.

There are also systems and proposed models based on ontologies. The proposal in [35] defined textual measures in addition to a traditional multidimensional model. In order to perform aggregations, the authors provided the AVG_KW function that used a simple ontology to aggregate keywords based on a predetermined distance and the LCA (Lowest Common Ancestor) of the ontology tree.

The study presented in [19] proposed a data model and some algebraic operations to analyze textual documents. In order to design a hierarchy for OLAP, the authors used ontologies and the hierarchies were classified into six types. The implementation stored hierarchies using preorder and postorder to execute searches and check containments.

The authors of [39] proposed a dynamic process (briefly explained in Section 4.2) to develop multidimensional schemas where ontologies were used as dimensions to provide semantic for OLAP.

The authors of [28] introduced a four-layer approach to multidimensional ontologies to represent the hierarchical and multidimensional concepts for OLAP.

The approach in [23] presented a multidimensional data model from semi-structured data for Twitter data warehouse. They developed a practical solution using reverse engineering and mining to discover facts, dimensions and hierarchies to construct the DW model in the ETL process. The authors also discussed the problem of maintenance of dynamic elements since new aggregations formed new hierarchies (see *Maintenance and Evolution of Data Warehouses*).

The study made in [14] proposed new steps to extract information from text documents at the ETL phase, maintaining relevant information of each document. The multidimensional model contains numeric measures and three kinds of dimensions: ordinary ones, such as a set of keywords; metadata dimensions, such as title, format, authors of the document; and category dimension, such as a hierarchy or an ontology. The next subsection briefly explains the data warehousing lifecycle with an emphasis on the ETL process.

Multimedia-centric:

There are several studies on the issue of integration of multimedia data in data warehouse environments, an emerging field called Multimedia Warehouse.

A multimedia warehouse called MediaHouse [48] is the oldest proposal we found that defined an approach to deal with multimedia data in data warehouses. The authors defined a fusion of star and snowflake models called the starflake model to handle the nature of multimedia objects. The similarity measure (symmetrical Tau) guided the integration of multiple features into the hierarchical.

The authors of [4] and [5] proposed a multidimensional and multiversion model to describe multimedia warehouse where multiple descriptors represented multimedia data and several computational functions described each descriptor. They implemented a medical application on acute myocardial infarction to demonstrate the multimedia approach by transforming electrocardiogram signals into content descriptors.

A structured model database that organizes the multimedia data in a multidimensional data cube based on XML is described in [41]. The approach described in [11] presented a generic UML model for complex data. The case study was based on images (towns and landscapes) and texts. Low-level characteristics of images (color, homogeneity, entropy) were submitted to a decision tree to classify the images as towns or landscapes. This label was automatically stored as a specific dimension. Furthermore, in the data integration phase, complex data were stored into XML documents using their own software.

In [46], the data model modified the standard multidimensional model including hierarchical structure to represent multimedia data. An image data model was provided to support a medical (pneumology) application, where each dimension, including the technical terms and the medical metadata was an XML file. Related technical and medical terms were hierarchically structured.

Visual Cube [21] is a cube model to support image-OLAP. The authors proposed a clustering structure measure to explore images, dynamic aggregation selection to improve computations and a new type of OLAP operation to support overlapping.

iCube [3] is a similarity-based data cube for medical images. It added a special dimension to store content-based features providing capabilities for OLAP similarity queries over images.

The approach in [25] presented a data structure that organizes video data on multidimensional data cubes with three types of dimensions (non-space, space-non space, space-space) and space measures.

B. Data Extraction, Cleansing, Transformation and Loading

The studies in this section consider unstructured (or semi-

structured) data integration in the ETL process.

The authors of [14] defined the lifecycle of a document warehousing system, with emphasis on textual ETL. They extended a methodology to transform traditional data warehouses into document warehouses focusing on the analysis of textual sources. The conceptual design phase defined the elements of the multidimensional model described in the previous subsection, while the implemented prototype focused on the ETL process. The case study was focused on Security and Prevention and the prototype used open source tools.

The authors of [38] described a methodology for building XML data warehouses using XQuery where the traditional ETL process was adapted for XML documents, creating intermediate XML documents, fact tables (star-schema) and links between dimensions and intermediate documents.

Besides the galaxy model, in [34] the authors presented a five stages design process to use document-centric XML documents for OLAP. The five stages were:

(1) user-requirements analysis;

(2) data sources analysis;

(3) confrontation;

(4) adaptation to user requirements or data sources enrichment; and

(5) the creation and load in multidimensional structures.

The authors of [24] also presented a method for XML data integration in which a unified tree structure was used to obtain a general overview of heterogeneous XML documents that should be loaded into a multidimensional document model.

The ETL process in [39] integrated operational data with unstructured document data which was indexed using semantic indexing, an information retrieval technique. The process indexed each document resulting on a set of terms, which were associated with the concepts of predefined ontologies, creating ontological trees for dimension hierarchies.

The study in [27] proposed an architecture for semantic annotation analysis focusing on the ETL process to extract facts, dimensions and hierarchies from semantic annotations based on domain ontologies. It was defined a multidimensional schema determined by ontology with semantic types of dimensions.

The authors of [23] used reverse engineering to construct a Twitter data model on which mining techniques such as clustering and sentiment analysis were applied to discover facts, dimensions and hierarchies to build a data warehouse for Twitter.

C. Data Warehouse Architecture

Baars et al. [6] and Alqarni et al. [2] described proposals for integrating structured and unstructured data. The first proposed a three-layer framework, where the logical layer analyzed structured data or unstructured content using OLAP or data mining. They discussed the proposed framework into different scenarios: Customer Relationship Management and Competitive Intelligence.

The second group of authors proposed a multilayer scheme for creating a relationship between structured data stored in a data warehouse and unstructured data, identifying related data. The connection between both types of data were represented using an XML schema.

The approach introduced in [2] described a methodology called X-Warehousing, which proposed a three-layer data schema, where the middle layer integrated schemata extracted from unstructured documents (bottom layer) and structured data warehouse (top layer). The physical level of the data warehouse was populated with XML documents based on the XML Schema. For this, the methodology defined a modeling mechanism to transform star and snowflake schema of existing data warehouses to XML Schema models.

Besides the model, an architecture for document warehouse was proposed in [43], which aimed to categorize the document base according to metadata or keywords (in the ETL phase). The authors applied the multidimensional modeling process to generate document cubes and the documents were stored as file pointers of the original documents at the document repository.

As part of the proposal described in [31] and [33], the proposed architecture of the contextualized warehouse associated facts between a corporate warehouse and a document warehouse based on the context of the documents.

A four-layer architecture was presented in the iNextCube [49]. The lower intermediate layer analyzed information networks to generate clusters and concept hierarchies. The layer above this integrated the TextCube and the TopicCube to provide information retrieval measures and latent semantic analysis and the top layer consisted on the user interface.

The authors of [46] also presented an architecture, which was structured into five blocks comprising ETL, warehouse, semantic metadata, processing and metadata maintenance and query processor blocks, and the warehouse block was composed of the dimension block and the fact block. They considered that some facts might consider other facts as dimensions. The query processor maps the query terms with the medical and technical terms and the necessary aggregations to resolve the query.

The work presented in [29] introduced an architecture, which dealt with two sources: structured data warehouse and text document warehouse. The Text OLAP module over the second DW integrated information retrieval, text mining, and information extraction technologies. The Consolidation OLAP module managed the process of consolidating relational OLAP and Text OLAP modules together.

D. Analytical Front-End Tools

OLAP models have been described previously, but there are two studies strictly related to results visualization for OLAP queries based on text [26] and [42].

The DocCube interface allowed users to manage the concept hierarchies (proposed as part of the DocCube model) and hierarchy levels [26]. DocCube included a 3D global

representation of documents and allowed OLAP operation analysis.

The approach presented in [42] described a method to dynamically generate an OLAP query from text interpretations and suggest charts to illustrate textual content. The architecture was divided in pre-processing and runtime components. In the pre-processing phase, a dictionary from the semantic layer was automatically generated, involving dimensions and measures. A workflow was constructed to build suggestions of queries at runtime. The authors provided a prototype, the Text-To-Query system (T2Q), to demonstrate the proposal.

Although the visualization was not the focus of the proposal, a tool for data navigation and visualization in the context of the medical prototype was implemented in [4].

E. Maintenance and Evolution of Data Warehouses

The extraction of information from unstructured data requires the maintenance and evolution of data warehouses.

The authors of [15] presented an overview of the state of the art of techniques for supporting keyword search in structured or semi-structured databases and commented on the challenges and opportunities of future research on unstructured data.

The challenges are:

1) diversity of data models;

2) queries that are highly expressive, but difficult to learn;

3) quality improvement on search (information retrieval was mentioned as an alternative); and

4) methods to evaluate and guide DW designs.

These issues are directly related to maintenance and evolution of data warehouses.

A survey on current researches on combining DWs with Web/XML data and technologies was provided in [32]. The authors discussed schema maintenance in the context of heterogeneity (federated and distributed DW architectures).

Additionally, in [45] it was discussed Semantic Web to represent Web content on data warehouses and semantic annotation as a useful resource for describing unstructured and semi-structured data. When working with unstructured data, the authors addressed the issue of domain ontologies to aid the multidimensional design process.

In the case of the Twitter data warehouse, the authors proposed an architecture for discovering dynamic elements such as dimension hierarchies in [23]. They dealt with the problem of maintenance of dynamic data that requires the adaptation of the DW by using the concept of slowly changing dimensions.

F. : Situational Business Intelligence

The studies in this section were explicitly separated from the previous discussions because they deal with an open research issue identified during the review: real-time Business Intelligence [45].

In order to make the process of decision-making more effective, in addition to stationary data in a DW, it is interesting to use transient data that decision makers are unaware. Recent studies have introduced the term *situational data*, which is data that have a close relationship with the domain specific problem and generally has a short lifetime, being interesting only for a small group of decision makers with a specific set of needs [1]. These data are related, for example, with the market, competitors or potential customers.

The ability to embed situational data in the decision process originated a new class of BI applications often labeled as situational BI, on-demand BI, or even collaborative BI. In [1], the authors introduced situational data to compose self-service BI, that is, a program that emphasizes the user's role as decision maker. The user can search, extract and group situational data through a continuous interaction with the application, without the need of any IT specialist intervention.

The proposed architecture in [1] included a new OLAP model called fusion cube, which can be dynamically created defining scheme, instances and metadata that can be associated with a set of annotations by the user. The idea was to allow the reuse of fusion cubes created by other users, since a query used data from traditional and/or fusion cubes.

A platform for situational awareness applications, SIE-OBI, was presented in [13]. The goal of SIE-OBI was to reduce time and effort extracting structured information from text streams (situational data) to place on the DW and query and analyze them in near real-time.

G. Qualitative Summary

In order to accomplish the qualitative synthesis proposed in section titled "*Qualitative criteria to evaluate primary studies*", the studies are summarized into the five categories defined. Each category is marked with an "X" when the topic is present in the study.

Some studies were spread in the more than one of the previous subsections because they present contributions on each one of them. In order to offer an overall description of the goals of each study, Table I presents the distribution of studies according to the subject matter. For example, we describe the study made in [25] as an academic research with case study (clearly evidencing implementation) that proposed a multidimensional model for multimedia data. We group studies when all meet the same criteria and have the same first author.

Because [32] and [45] are review articles, they did not present new proposals and were not included in Table I. The survey made in [32] is an academic work that exposed the use of XML for DW from three angles: technology integration, extensions of DW and OLAP for XML Web data, and the combination of OLAP and information retrieval to improve applications in order to extend OLAP to support unstructured document analysis.

On the other hand, the work presented in [45] is also academic and examined current work and open issues in spatio-temporal data warehouses, real-time data warehousing and Semantic Web Data Warehousing and OLAP.

TABLE I
QUALITATIVE EVALUATION SUMMARY I

Studies	Classes					Doc.-centric	MM-centric	Evidence of implemen-tation	Case study	Provenance
	A	B	C	D	E					
[2]			X			X		Yes	Yes	Academic
[4]; [5]		X		X			X	Yes	Yes	Academic
[6]			X			X		No	No	Academic
[7]	X					X		Yes	No	Industry
[8]	X					X		No	No	Industry
[11]	X	X				X	X	Yes	Yes	Academic
[14]	X	X	X			X		Yes	No	Industry
[16]	X					X		Yes	No	Academic
[22]	X					X		Yes	Yes	Academic
[50]	X					X		Yes	Yes	Academic + Industry
[19]	X					X		Yes	Yes	Industry
[20]	X					X		Yes	Yes	Academic
[25]	X						X	Yes	Yes	Academic
[26]	X			X		X		Yes	No	Academic
[29]		X	X			X		No	No	Academic
[31]; [33]	X		X			X		Yes	Yes	Academic
[34]	X	X				X		Yes	No	Academic
[36]; [35]; [37]	X					X		Yes	No	Academic
[38]	X	X				X		Yes	Yes	Academic
[39]	X	X				X		Yes	No	Academic + Industry
[41]	X						X	Yes	No	Academic
[42]				X		X		Yes	Yes	Industry
[43]	X					X		Yes	Yes	Academic
[44]	X		X			X		Yes	Yes	Academic
[46]	X	X	X		X		X	Yes	Yes	Academic
[48]	X		X				X	Yes	Yes	Academic
[24]	X	X				X		Yes	No	Academic
[23]	X	X	X		X	X		Yes	Yes	Academic
[1]	X	X	X		X	X		No	No	Academic
[13]		X	X			X		Yes	Yes	Industry
[28]	X		X			X		Yes	Yes	Academic
[27]	X	X	X			X		Yes	Yes	Academic
[21]	X						X	Yes	Yes	Academic + Industry
[49]	X		X			X		Yes	No	Academic
[3]	X						X	Yes	Yes	Academic
[9]	X					X		Yes	No	Academic
Total	33	14	14	4	3	32	9	36+	23+	34 Ac.
(%)	**82.5**	35	35	**10**	**7.5**	**80**	22.5	90	57.5	85

The percentages in Table I were calculated as the number of studies matching the criterion divided by the total number of studies, which in the case is 40. The topic *Multidimensional modeling and Data warehouse design* represents 82.5% of the subjects addressed in the studies. This high number could indicate that most of the studies are concentrating their efforts on the capture of the nature of the unstructured data in order to extend the multidimensional model and OLAP operators to manipulate them. Furthermore, 52.5% of the studies (21 studies) addressed more than one subject in their proposals, while 45% of the contributions (18/21 - 85.7%) present solutions integrating multidimensional modeling for non-traditional data in addition to an architecture and/or data integration.

It was not possible to identify a standard approach to merge unstructured data in data warehouse environments, since there are different proposals, which bring together unstructured data and traditional warehouses, proposed data models for document data and data models for multimedia data. The common feature of all proposals is the integration of semantics in multi-dimensional analysis. Moreover, it was not possible to identify the existence of a standard procedure to develop the data warehousing process since most of the approaches captured specific situations and used different data integration strategies, extraction techniques and models.

At this point, the challenges raised to continue the work in this field are the diversity of data models, OLAP models, front-end tools, maintenance and strategies of evolution to provide data quality and diversity of procedures to guide system design.

Most of the studies (80%) introduced proposals of textual data, but 22.5% of the publications covered multimedia data, mainly images. Most recent efforts to integrate unstructured or semi-structured data into data warehouses are being developed for document or text warehouses. Nevertheless, it is important to raise the attention to emerging application domains in business intelligence such as multimedia because new domains require new OLAP domains [45]. One implementation studied covered both document and multimedia data types as complex objects. It is worth to mention that the majority of the contributions related to DW architecture took into consideration various types of unstructured data, but in terms of implementation, they focus only on one type.

OLAP extensions for textual data used mainly five technologies: Text Mining, Information Retrieval and Extraction, Ontology and Semantic Web, the latter being selected to represent Web content and semantic relations and, thus, a path to solve maintenance and evolution issues, by using RDF, ontologies (and related approaches).

During the development of this work, we observe that many recent solutions are strongly investing into data integration (such as [30], [12]) to exploit the already quite stable and robust traditional Business Intelligence framework.

All publication provenances were established based on the authors' affiliation. The provenance can be seen as a relevant qualitative information as it has an influence on experiments, patents and research investments. The vast majority of the studies were developed in the academic realm (85% - counting joint efforts). Just 22.5 percent (also counting overlaps) of the selected studies were industry experiences and no software patent was found.

According to Table I, 90% of studies had evidence of implementation and 57.5% explicitly reported an empirical research with implementation and experimental results. Since the field is still new, there is a shortage of case studies. Investments on research will bring more experimental results.

We believe that the presence of business or industry branch in research will raise the number of experimental results and software patents in the field.

IV. Conclusion

In this survey, we classified and discussed the current issues related to unstructured data in the context of data warehouses and OLAP. The systematic review is a very careful process. Once the protocol was defined, the authors had to follow a very detailed and cumbersome procedure to achieve its goal. In order to make the systematic review repeatable and easier to update, the StArt tool supported our work.

An important aspect to remember is that not all approaches for unstructured data warehouse were available for reading (most conference papers), thus, they were not included in the review.

Through the adopted systematic review, it was possible to identify and summarize a set of studies, extracting relevant information to establish the state of the art and answer the research question. Then, the analysis of the selected articles revealed the existence of initiatives for integrating textual and multimedia sources in data warehouse environments. Moreover, it was possible to identify specific research fields, which concentrate the research.

A deep analysis of the literature revealed that the research community does not agree on a solution to integrate unstructured data in data warehousing environment. Solutions reached mostly multidimensional representations of these types of data. Some interesting solutions were proposed but no shared framework has been devised yet. We observe that there were solutions to text and image warehouses to create effective query cubes based on the unstructured nature of data sources and the introduction of semantics. Recent approaches offered some support to this kind of data but not robust implementation to treat unstructured data as a native kind of data in OLAP.

Based on the ongoing researches and the summarization board in the previous section, we identified some major concerns:

- to build OLAP queries combining unstructured and traditional data;
- to enhance multimedia OLAP;
- to reduce the response time of systems, enabling real time information;
- to improve query interfaces and visualization for unstructured data to convey insights to end-users;
- to develop semantic understanding for metadata and schemata to support maintenance and evolution;
- to add situational data in real-time (or near); and
- to discover data relations from data sources to define multidimensional schemata, facts, dimensions, hierarchies and aggregations.

The above research aspects require further investigation. There are still open issues and plenty of quality research to do in the field of this review.

Considering the maturity of data warehousing systems, there will be more demand for advanced and intelligent support to integrate unstructured data in Data Warehousing. Thus, it is essential that both industry and academic researchers be ready to deliver effective solutions that could be deemed acceptable by the market.

Acknowledgement

This research is part of a funding support granted by FAPESP (São Paulo Research Foundation), Process number 2011/12115-1.

References

[1] Abello, A., Darmont, J., Etcheverry, L., Golfarelli, M., Mazón, J., Naumann, F., Vossen, J.. “Fusion Cubes: Towards Self-Service Business Intelligence”. International Journal of Data Warehousing and Mining, 9(2), 66-88, 2013.

[2] Alqarni, A. A., & Pardede, E.. Integration of data warehouse and unstructured business documents. Proceedings of the 15th International Conference on Network-Based Information Systems (NBIS), 2012.

[3] Annibal, L., Felipe, J., Ciferri, C., & Ciferri, R.. iCube: A similarity-based data cube for medical images. Proceedings of the 23rd IEEE International Symposium on Computer-Based Medical Systems (CBMS), (pp. 321-326), 2010.

[4] Arigon, A.-M., Miquel, M., & Tchounikine, A.. Multimedia data warehouses: A multiversion model and a medical application. Multimedia Tools and Applications, 35(1), 91-108, 2007.

[5] Arigon, A.-M., Tchounikine, A., & Miquel, M.. Handling multiple points of view in a multimedia data warehouse. ACM Transactions on Multimedia Computing, Communications and Applications, 2(3), 199-218, 2006.

[6] Baars, H., & Kemper, H.-G.. Management support with structured and unstructured data - An integrated business intelligence framework. Information Systems Management, 25(2), 132-148, 2010.

[7] Barczynski, W. M., Brauer, F., Mocan, A., Schramm, M., & Froemberg, J.. BI-style relation discovery among entities in text. Proceedings of International Conference on Data Engineering, 2010.

[8] Bordawekar, R. R., & Lang, C. A.. Analytical processing of XML documents: Opportunities and challenges. SIGMOD Record, 34(2), 27-32, 2005.

[9] Boukrâ, D., Boussaïd, O., & Bentayeb, F.. OLAP Operators for Complex Object Data Cubes. In Advances in Databases and Information Systems (Vol. 6295, pp. 103-116). Springer Berlin Heidelberg, 2010.

[10] Boussaid, O., & Boukraâ, D.. Multidimensional Modeling of Complex Data. In Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1358-1364). IGI Global, 2009.

[11] Boussaid, O., Tanasescu, A., Bentayeb, F., & Darmont, J.. Integration and dimensional modeling approaches for complex data warehousing. Journal of Global Optimization, 37(4), 571-591, 2007

[12] Carrasco, R. A., Muñoz-Leiva, F., & Hornos, M. J.. A multidimensional data model using the fuzzy model based on the semantic translation. Information Systems Frontiers, 15(3), 351-370, 2013.

[13] Castellanos, M., Chetan, G., Wang, S., Dayal, U., & Durazo, M.. A platform for situational awareness in operational BI. Decision Support Systems, 52, 869-883, 2012.

[14] Cembalo, A., Pisano, F. M., & Romano, G.. An approach to document warehousing system lifecyle from textual ETL to multidimensional queries: A proof-of-concept prototype. Proceedings of the 6th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS), 2012.

[15] Chen, Y., Wang, W., Liu, Z., & Lin, X.. Keyword search on structured and semi-structured data. Proceedings of International Conference on Management of Data and 28th Symposium on Principles of Database Systems (SIGMOD-PODS), 2009.

[16] Chen, Z., Garcia-Alvarado, C., & Ordoñez, C.. Enhancing document exploration with OLAP. Proceedings of IEEE International Conference on Data Mining (ICDM), 2010.

[17] Fabbri, S., Hernandes, E. M., Thommazo, A. D., Belgamo, A., Zamboni, A., & Silva, C.. Managing Literature Reviews Information through Visualization. Em L. A. Maciaszek, A. Cuzzocrea, & J. Cordeiro (Ed.), Proceedings of International Conference on Enterprise Information Systems (ICEIS) (pp. 36-45). SciTePress, 2012.

[18] Han, J., & Kamber, M.. Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann, 2006.

[19] Inokuchi, A., & Takeda, K.. A method for online analytical processing of text data. Proceedings of International Conference on Information and Knowledge Management, 2007.

[20] Janet, B., & Reddy, A. V.. Cube index for unstructured text analysis and mining. Em ACM International Conference Proceeding Series, 2011.

[21] Jin, X., Han, J., Cao, L., Luo, J., Ding, B., & Lin, C. X.. Visual cube and on-line analytical processing of images. Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM) (pp. 849-858). ACM, 2010.

[22] Lin, C., Ding, B., Han, J., Zhu, F., & Zhao, B.. "Text Cube: Computing IR Measures for Multidimensional Text Database Analysis. Proceedings of 8th IEEE International Conference on Data Mining (ICDM), (pp. 905-910), 2008.

[23] Mansmann, S., Rehman, N. U., Weiler, A. S., & H., M.. "Discovering OLAP dimensions in semi-structured data". Proceedings of International Conference on Information and Knowledge Management, (pp. 9-16), 2012.

[24] Messaoud, B., I., F., J., Z., & G.. "A first step for building a document warehouse: Unification of XML documents". Proceedings of International Conference on Research Challenges in Information Science, 2012.

[25] Mianshu, C., Yu, S., Zhonghua, S., Hexin, C., & Aijun, S.. "Multimedia database retrieval based on data cube". Proceedings of International Conference on Audio, Language and Image Processing (ICALIP), 2008.

[26] Mothe, J., Chrisment, C., Dousset, B., & Alaux, J.. "DocCube: Multi-dimensional visualisation and exploration of large document sets". Journal of the American Society for Information Science and Technology, 54(7), 650-659, 2003.

[27] Nebot, V., & Berlanga, R.. "Building data warehouses with semantic web data". Decision Support Systems, 52(4), 853–868, 2012.

[28] Neumayr, B., Anderlik, S., & Schrefl, M.. "Towards Ontology-based OLAP: Datalog-based reasoning over multidimensional ontologies". Proceedings of the 15th International Workshop on Data warehousing and OLAP, (pp. 41-48), 2012.

[29] Park, B.-K., & Song, I.-Y.. "Toward total business intelligence incorporating structured and unstructured data". ACM International Conference Proceeding Series, 2011.

[30] Pedersen, T., Pedersen, D., & Riis, K.. "On-demand multidimensional data integration: toward a semantic foundation for cloud intelligence". The Journal of Supercomputing, 65(1), 217-257, 2013.

[31] Pérez, J. M., Berlanga, R., & Aramburu, M. J.. "A relevance model for a data warehouse contextualized with documents". Information Processing and Management, 45(3), 356-367, 2009.

[32] Pérez, J. M., Berlanga, R., Aramburu, M. J., & Pedersen, T. B.. "Integrating data warehouses with Web data: A survey. IEEE Transactions on Knowledge and Data Engineering", 20(7), 940-955, 2008.

[33] Pérez-Martínez, J. M., Berlanga-Llavori, R., Aramburu-Cabo, M. J., & Pedersen, T. B.. "Contextualizing data warehouses with documents". Decision Support Systems, 45(1), 77-94, 2008.

[34] Pujolle, G., Ravat, F., Teste, O., Tournier, R., & Zurfluh, G.. "Multidimensional database design from document-centric XML documents". Lecture Notes in Computer Science, 2011;

[35] Ravat, F., Teste, O., & Tournier, R.. "OLAP aggregation function for textual data warehouse". Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS), 2007.

[36] Ravat, F., Teste, O., Tournier, R., & Zurfluh, G.. "Top_keyword: An aggregation function for textual document OLAP". Lecture Notes in Computer Science, 2008.

[37] Ravat, F., Teste, O., Tournier, R., & Zurlfluh, G.. "A conceptual model for multidimensional analysis of documents". Lecture Notes in Computer Science, 2007.

[38] Rusu, L. I., Rahayu, W., & Taniar, D.. "On building XML data warehouses". Lecture Notes in Computer Science, 2004.

[39] Sciarrone, F., Starace, P., & Federici, T.. "A business intelligence process to support information retrieval in an ontology-based environment". Proceedings of the 9th International Conference on Intelligent Systems Design and Applications (ISDA), 2009.

[40] Staples, M., & Niazi, M.. "Experiences using systematic review guidelines". Journal of Systems and Software, 80(9), 1425-1437, 2007.

[41] Sun, Y., Chen, H., Chen, M., Wang, X., & Sang, A.. "Multi-dimension multimedia retrieval model implementation based on XML database". Proceedings of International Conference on Signal Processing Systems (ICSPS), 2009.

[42] Thollot, R., Brauer, F., Barczynski, W. M., & Aufaure, M.-A.. "Text-to-query: Dynamically building structured analytics to illustrate textual content". ACM International Conference Proceeding Series, 2010.

[43] Tseng, F. S., & Chou, A. Y.. "The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence". Decision Support Systems, 42, 727-744, 2006.

[44] Tseng, F. S., & Lin, W.-P. D.-t.. "D-tree: a multi-dimensional indexing structure for constructing document warehouses". Journal of Information Science and Engineering, 22, 819-841, 2006.

[45] Vaisman, A., & Zimanyi, E. D.. "Data warehouses: Next challenges". Lecture Notes in Business Information Processing, 2012.

[46] Vanea, A., & Potolea, R.. "A hierarchical semantically enhanced multimedia data warehouse". Proceedings of the 6th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), 2010.

[47] William H. Inmon, A. N.. "Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence". Prentice Hall, 2007.

[48] You, J., Dillon, T., Liu, J., & Pissaloux, E.. "On hierarchical multimedia information retrieval" Proceedings of IEEE International Conference on Image Processing, 2001.

[49] Yu, Y., Lin, C. X., Sun, Y., Chen, C., Han, J., Liao, B., Zhao, B.. "iNextCube: information network-enhanced text cube". Proceedings of VLDB Endow., 2(2), 1622-1625, 2009.

[50] Zhang, D., Zhai, C., Han, J., Srivastava, A., & Oza, N. (2009). "Topic modeling for OLAP on multidimensional text databases: Topic cube and its applications". Statistical Analysis and Data Mining, 2(5-6), 378-395.

Sahudy Montenegro González received the BS degree in Computer Science from the Universidad de La Habana in 1994, MS degree and PhD degree in Electrical Engineering from School of Electrical and Computing Engineering at State University of Campinas. She is a full professor of Computer Science at Federal University of São Carlos, Brazil. Her research interests include multidimensional databases, OLAP, data warehousing, new database technologies, text mining and information and multimedia retrieval systems.

Talita dos Reis Lopes Berbel received the BS degree in Data Processing from the Faculdade de Tecnologia de Sorocaba in 2006 and received the degree in Computer Engineer from Faculdade de Engenharia de Sorocaba in 2010. She completed the MBA course in Project Management and Business in the Faculdade de Engenharia de Sorocaba in 2014. Currently, she is a Masters student in Computer Science at Federal University of São Carlos. She works as an associate professor of Computer Engineering and has experience on databases, system analysis and software engineering. Her current research areas include multidimensional databases, text mining and OLAP.