Decision Support Systems 72 (2015) 44–59

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Multidimensional analysis model for a document warehouse that includes textual measures

Martha Mendoza [a,b,*], Erwin Alegría [a], Manuel Maca [a], Carlos Cobos [a,b], Elizabeth León [c]

[a] Information Technology Research Group (GTI), Universidad del Cauca, FIET 422 Popayán, Colombia
[b] Electronic and Telecommunications Engineering Faculty, Universidad del Cauca, Colombia
[c] Engineering Faculty, Universidad Nacional de Colombia, Colombia

ARTICLE INFO

Article history:
Received 15 October 2012
Received in revised form 6 February 2015
Accepted 8 February 2015
Available online 16 February 2015

Keywords:
Document warehouse
OLAP
Textual measures
Text warehouse

ABSTRACT

Data warehouses and On-Line Analytical Processing tools, OLAP, together permit a multi-dimensional analysis of structured data information. However, as business systems are increasingly required to handle substantial quantities of unstructured textual information, the need arises for an effective and similar means of analysis. To manage unstructured text data stored in data warehouses, a new multi-dimensional analysis model is proposed that includes textual measures as well as a topic hierarchy. In this model, the textual measures that associate the topics with the text documents are generated by Probabilistic Latent Semantic Analysis, while the hierarchy is created automatically using a clustering algorithm. Documents are then able to be queried using OLAP tools. The model was evaluated from two viewpoints — query execution time and user satisfaction. Evaluation of execution time was carried out on scientific articles using two query types and user satisfaction (with query time and ease of use) using statistical frequency and multivariate analyses. Encouraging observations included that as the number of documents increases, query time increases as a lineal, rather than exponential tendency. In addition, the model gained an increasing acceptance with use, while the visualization of the model was also well received by users.

1. Introduction

Data warehouses are widely used today for analyzing large amounts of data in a variety of contexts (e.g. marketing, health, education, and research [29]). The tool most commonly used for querying available information in data warehouses is On-Line Analytical Processing (OLAP), due to its ease of use and capacity to conduct flexible analysis of the data in real time [23,25]. It has been routinely applied in many different domains [9,12,14,21,22,26,31].

The amount of unstructured data has risen significantly due to growth of the internet, more document information in business systems, and observations and comments added to object-relational databases. It is thought that no more than 20% of information extracted from data warehouses relates to simple numeric data, i.e. the remainder is hidden in non-numeric data, or documents [27]. Unfortunately, conventional OLAP tools do not allow adequate management of this information in data cube queries. In the area of health, for example, information relating to the medical treatment of a patient usually comprises *name of patient*, *prescribed medication*, *appointment date*, *diagnostic*, and *observations*. This latter field, the *observations* attribute, is unstructured and in modeling a data warehouse is usually omitted or, in the best case scenario, modeled as a descriptor. Regardless, it cannot be used in OLAP operations (drill-down, roll-up, slice, dice, and pivot). As a result, business analysts are often making decisions while omitting this important information stored in texts.

In the research field, to inquire about the state of the art of a particular subject, it is necessary to engage a great amount of documents, some of which fail to contribute in the desired way. Were it possible to store a scientific article with its percentage of relation to certain areas or subjects, this would allow researchers to review articles closest to the subject about which they are inquiring. The handling of unstructured text opens the possibility of conducting a more complete analysis of the information existing in organizations.

A key challenge that data warehouses and OLAP tools face is thus to incorporate unstructured data and textual measures so they are navigable in the warehouse [5]. This paper proposes a multidimensional model that incorporates textual measures — generated with Probabilistic Latent Semantic Analysis (PLSA) — in combination with an automatically-created topic hierarchy. The model thus allows documents to be queried using OLAP operations. In addition, a prototype of OLAP is put forward that permits querying a set of scientific articles stored in a warehouse.

The paper is organized as follows: Section 2 presents the most relevant related works. The architecture of the proposed model is outlined in Section 3. Section 4 explains the technical aspects of the model. Section 5 discusses in detail the creation of textual measures and the

* Corresponding author at: Information Technology Research Group (GTI), Universidad del Cauca, Sector Tulcán Office 450, Popayán, Colombia. Tel.: +57 28366524; fax: +57 28209810.

http://dx.doi.org/10.1016/j.dss.2015.02.008
0167-9236/© 2015 Elsevier B.V. All rights reserved.

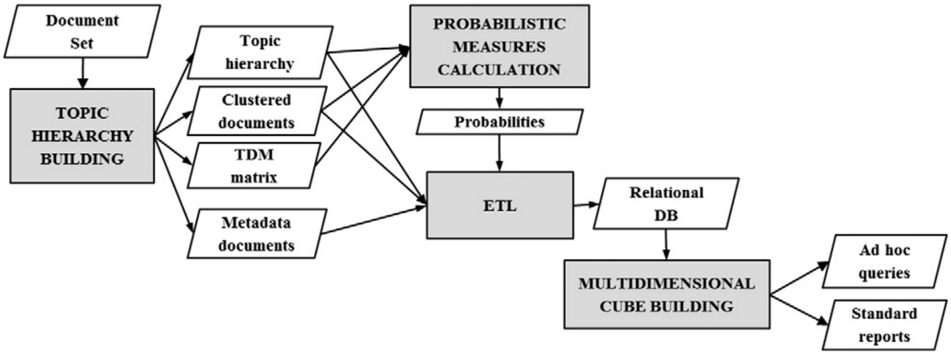

Fig. 1. General architecture of the proposed model.

aggregation function for each of them. Section 6 presents the automatic creation of the topic-hierarchy using an evolutionary clustering algorithm, while Section 7 explains how to obtain the probabilities needed to relate the documents to the topics. The results of the evaluation, based on textual measures are found in Section 8. Section 9 presents visualization of queries conducted in the document warehouse using an OLAP tool. Section 10 presents conclusions and future work.

2. Related work

The research conducted on text analysis using OLAP tools can be divided in two categories [30]: text as categorical data, and text as an OLAP component.

In the approach that treats text as categorical data, document classification methods are applied to class labels within a category. This model allows drill-down or roll-up operations on the category dimensions, but uses a traditional multidimensional cube (only with structured data). In addition, the classification process in these proposals requires training data (documents previously assigned to categories) and this information is not available in every scenario. There are several representative works for this approach. Cody et al. [3] from IBM in 2002 proposed some general ideas of how to integrate text mining algorithms with OLAP. Nonetheless, in this work the cubes are still the traditional ones and with a low integration level (at an external level) of both technologies. In addition, the subject of efficient materialization of the cubes that handles the text dimension is not addressed. There is also McCabe's work [19] from 2000, focused on information retrieval, which presents the terms occurrence in the document as a fact for the multidimensional model, with dimensions for the term and document; hierarchies for time and localization; and a category dimension that contains a subject category for the term. The measure is the weight of the term within the document (term frequency). The main objective of this research consists in conducting searches by terms within the document, combined with time and location. In this respect, Tseng and Chou [27] presented in 2006 a general framework for a document warehouse (data sources, pre-processing of documents and data presentation). They define a dimensional model for this type of warehouse. The dimensions are classified in three categories: ordinary dimension, with the set of keywords that allows users to locate the desired documents directly; a metadata dimension that gives the document information (e.g. title, author, publisher, date, etc.); and finally the category dimension. This last dimension can be either a WordNet-based hierarchy, or one defined by the user. The fact table is formed by a keyword composed of the foreign keys of the dimensions previously named, attributes used to derive the document measures (term frequency), and a Document ID column that represents the document identifier (a foreign key from a dimension that holds all the documents identifiers and the file path). In this model the table granularity is by keyword, which makes user navigation through the document very complex. In addition it only contemplates counting measures, not a probabilistic measure that allows giving a better approximation and relevance of the subjects addressed by the document.

In turn, Franck Ravat et al. [8] propose a multidimensional model where they integrate an aggregation function, with the advantage of combining qualitative and quantitative analysis, i.e. the keyword analysis from a publication, with the objective of providing a view of the content of the publication. They provide a function that aggregates a set of keywords into a smaller set of more general keywords. This process is based on a conceptual model that provides: concepts or subjects adapted for supporting no-numeric textual measure as a hierarchical representation of the concepts analyzed with the help of ontologies, and a new concept of OLAP textual aggregation processing unit with the use of domain ontologies. This model requires ontology management, which makes it more complex since it corresponds to a concept domain hierarchy where each node represents a concept and each edge between nodes models an "is a" relationship. The model proposed in this paper introduces measures of the document content different from the traditional TF and IDF. The aggregation function of these textual measures is created by means of MultiDimensional eXpressions (MDX) and traditional programming language, besides the granularity of the fact table is at document level and not at a term level (keywords) as stated in these works, which allows greater query power on the data.

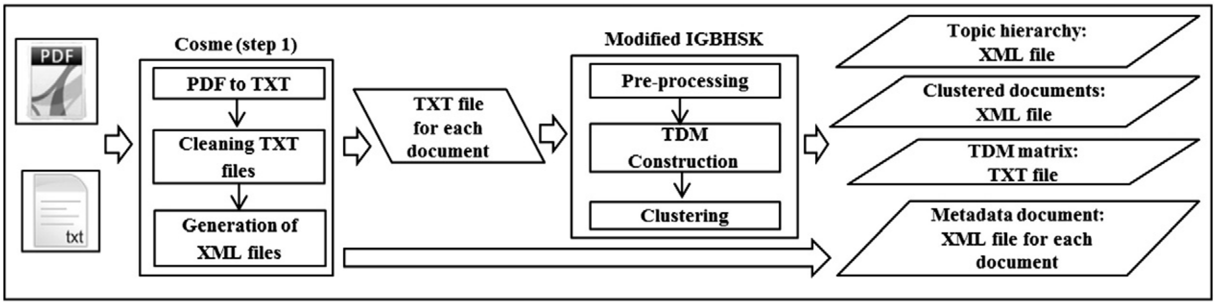

Fig. 2. Topic Hierarchy Building.