Expert Systems With Applications 93 (2018) 118–133

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Building a contextual dimension for OLAP using textual data from social networks

Karel Gutiérrez-Batista, Jesús R. Campaña*, Maria-Amparo Vila, Maria J. Martin-Bautista

Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

ARTICLE INFO

Article history:
Received 15 May 2017
Revised 19 September 2017
Accepted 4 October 2017
Available online 5 October 2017

Keywords:
OLAP
Data warehousing
Multidimensional model
Context detection
Text clustering

ABSTRACT

Due to the continuous growth of social networks the textual information available has increased exponentially. Data warehouses (DW) and online analytical processing (OLAP) are some of the established technologies to process and analyze structured data. However, one of their main limitations is the lack of automatic processing and analysis of unstructured data (specifically, textual data), and its integration with structured data. This paper proposes the creation, integration and implementation of a new dimension called *Contextual Dimension* from texts obtained from social networks into a multidimensional model. Such a dimension is automatically created after applying hierarchical clustering algorithms and is fully independent from the language of the texts. This dimension allows the inclusion of multidimensional analysis of texts using contexts and topics integrated with conventional dimensions into business decisions. The experiments were carried out by means of a freeware OLAP system (Wonder 3.0) using real data from social networks.

1. Introduction

The popularity and dramatic increase of the use of social networks in the last ten years has led to the creation of huge amounts of textual data generated by tens of millions of users on a daily basis (Guille, Hacid, Favre, & Zighed, 2013). The automatic massive processing and analyzing capabilities of most current technologies and systems are not enough to deal with such a huge quantity of heterogeneous, semi-structured, and unstructured data. Another challenge is the integration of textual information with traditional data, so that organizations can use this new resource and the possibilities offered by social network data. Content in social media messages can be very relevant to queries and decision makers need to use this content in order to take into account the full extent of the information available.

The processing of massive data involves the summarization and clustering of the data. For such a purpose, DW and OLAP are presented as the most adequate technologies, as they base their success on the advantages of integration, storage, and the operations of a multidimensional model. DW and OLAP thus allow the development of aggregations through conventional and unconventional dimensions for heterogeneous data. For the specific case of textual data, these systems need to undergo some kind of transformation to bring data to a more structured format, thus facilitating the analysis. In order to apply DW and OLAP to analyze the textual information provided by social networks, it is necessary to detect the main terms of the domains for the contexts of the texts. This would allow the decision makers to segment each context found and to treat it by taking advantage of the features and capabilities provided by multidimensional analysis.

There are many papers devoted to the use of the DW and OLAP technologies for the study of textual data present in the different social networks. Most of these papers involve information retrieval, sentiment analysis, recommendation systems, etc. These approaches do not take into account the context to which textual data actually belong, or use a predefined context.

In a previous study a new representation for textual data and their associated operations for query definition were presented (Martin-Bautista, Martínez-Folgoso, & Vila, 2015). Based on that representation, in Martin-Bautista, Molina, Tejeda-Avila, and Vila (2013) a new definition and implementation (Wonder 3.0) of a textual hierarchy (AP-dimension) is presented. Wonder is an OLAP system based on PostgreSQL that gives support to textual dimensions. We will use the definition of textual dimension and Wonder 3.0 to manage the domain hierarchy included in our proposal which is part of the context hierarchy. We have also previously carried out research about context detection and the influence of sentiment words in the process of context detection, which have been used in this proposal.

* Corresponding author.
E-mail addresses: karel@decsai.ugr.es (K. Gutiérrez-Batista), jesuscg@decsai.ugr.es (J.R. Campaña), vila@decsai.ugr.es (M.-A. Vila), mbautis@decsai.ugr.es (M.J. Martin-Bautista).

https://doi.org/10.1016/j.eswa.2017.10.012
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

The findings from our previous research have been integrated into a new methodology which uses data from social networks and builds a contextual dimension. This contextual hierarchy of textual dimensions contains the semantics associated to texts, which allows the business analysts to implement a detailed study by means of the use of an OLAP system (Wonder 3.0) on the topics discussed by users in social networks.

This study offers a novel solution to the current difficulty of achieving the application of multidimensional analysis on heterogeneous data by integrating textual data from social networks. It also improves results from the automatic detection of the contexts included in texts.

The contributions of this paper are:

- A new methodology for the creation of a data warehouse with textual dimension organized by contexts (set of topics) named Contextual Dimension, and its implementation in a real OLAP system (Wonder 3.0). This dimension is created automatically using a system to process and organize data from social networks. The methodology uses tools such as Multilingual Central Repository 3.0 (MCR 3.0) in order to make the process multilingual. The Contextual Dimension is a data warehouse dimension extracted from texts formed of two components: a context hierarchy composed of groups of topics discussed in a text, and within each level and context of this hierarchy, a domain hierarchy including the terms included in the texts.
- The Contextual Dimension will allow the decision makers to analyze the data from social networks selecting a context that has been automatically extracted and organized from social network data. The originality of our proposal lies in the fact that decision makers do not need to know contexts in advance to perform queries. In this analysis, it is possible to combine textual information with other traditional attributes, e.g. time of the day, and day of the week when the comments are made.
- The procedure to perform the integration is automatic and independent from the language in which the text is written. This enables the analysis of textual data in social networks using aggregations involving conventional and unconventional dimensions on heterogeneous data.

The rest of the article is organized as follows: Section 2 summarizes the main work developed in this research area. Section 3 explains the structure of the proposed dimension and presents the formal definitions supporting such a structure. Section 4 presents the creation process of the contextual dimension and its integration into a multidimensional model from the implementation point of view. The experimental results of our proposal and its discussion are shown in Section 5. Finally, Section 6 summarizes the conclusions which can be obtained from the implemented work and introduces some ideas regarding future research where we will further study the topics dealt within this research.

2. Related work

As previously mentioned, the main idea of our research is the automatic creation and integration of a contextual dimension into a multidimensional model, enabling users, organizations, and researchers to analyze data from social networks according to the main detected contexts and topics. Due to this reason, we study and analyze the main contributions related to the detection of contexts from texts in social networks, and the multidimensional analysis on heterogeneous data integrating both textual data and conventional data in social networks.

This section consists of two parts; the first describes relevant studies for context detection and the second presents some significant studies related to multidimensional analysis in social networks. In each case, the main characteristics of our approach which make it different from others are highlighted.

2.1. Context detection

Allan et al., in the first study on topic detection and tracking (Allan, Carbonell, & Doddington, 1998a; Allan, Papka, & Lavrenko, 1998b), focus on the exploration of techniques for the detection of the appearance of new topics and for tracking their reappearance and evolution. In Young-Woo and Sycara (2004), the authors mention the fact that topic extraction must deal with untagged data and that new subsets of events with similar contents are clustered together, therefore clustering algorithms are a good choice for discovering unknown events.

This latter approach is the closest to the contribution of this study, as it adjusts to our task of automatically detecting contexts from large volumes of textual data to achieve their integration as a new dimension in a multidimensional model. In order to achieve this goal, we will specifically focus on the application of hierarchical clustering methods.

There are many papers on the automatic categorization of texts, such as Chung-Hong (2012), where a mechanism which allows the clustering in real time of the content of the Twitter microblogs to detect events is established. Then, to assess the relationship amongst the different events, a method which combines the advantages of a clustering algorithm and a supervised learning model is developed. Skarmeta et al. present a study of the use of a semi-supervised agglomerative hierarchical clustering (ssAHC) algorithm, which allocates the texts to predefined categories (Skarmeta, Bensaid, & Tazi, 2000). Zheng and Li (2011) propose a new approach for the semi-supervised hierarchical clustering based on the establishment of a connection between the dendrogram of the hierarchical clustering algorithm and the ultrametric distance matrix, where the restrictions have been introduced by means of the *"Triple-wise relative constraints"* method. The previous studies share the condition that the set of tags or categories in which the texts will be grouped is known beforehand.

Hierarchical clustering algorithms have been widely studied in the literature for different kind of situations, including textual data (Deshmukh, Kamble, & Dandekar, 2013; RaghavaRao, Sravankumar, & Madhu, 2012). In Voorhees (1986) and Willett (1988) a general overview of the types of traditional agglomerative hierarchical clustering algorithms and their operation in the context of textual data is presented.

Gao, Gao, He, Wang, and Sun (2013) propose a new topic detection algorithm from pieces of news published on the internet on large disasters, based on group average hierarchical clustering (GAHC). The main idea of such an algorithm consists of dividing large data into smaller clusters and then using hierarchical clustering on these groups to generate the final topics. A practical tool for helping journalists and news readers to find interesting topics in message streams without feeling overwhelmed is presented in Martin, Corney, and Goker (2013). In this case, a time dependent variation of the classical tf-idf approach is presented. Sentences are grouped into bursts, often appearing in the same messages, so as to identify the emerging topics in the same time window. The experiments were carried out with Twitter data related to sports and politics. The study in Xiaohui, Xiaofeng, Yunming, Shengchun, and Xutao (2013) shows a conceptual graph containing concepts as nodes. Nodes connected by an edge share the same topic terms. When the hierarchical clustering is executed in this conceptual graph, the behavior curves for highly correlated concepts are grouped as topics.

It is worth mentioning that these studies are mainly oriented to topic or event detection in social networks, where the data belong to a specific domain and happen during a given interval of time.