Information Systems 53 (2015) 87–106

Contents lists available at ScienceDirect

Information Systems

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

Advanced topic modeling for social business intelligence

Enrico Gallinucci, Matteo Golfarelli, Stefano Rizzi *

DISI – University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy

ARTICLE INFO

Article history:
Received 10 March 2014
Received in revised form
21 January 2015
Accepted 15 April 2015
Available online 28 April 2015

Keywords:
Business intelligence
Social media
User-generated content
Multidimensional modeling

ABSTRACT

Social business intelligence combines corporate data with user-generated content (UGC) to make decision-makers aware of the trends perceived from the environment. A key role in the analysis of textual UGC is played by topics, meant as specific concepts of interest within a subject area. To enable aggregations of topics at different levels, a topic hierarchy has to be defined. Some attempts have been made to address the peculiarities of topic hierarchies, but no comprehensive solution has been found so far. The approach we propose to model topic hierarchies in ROLAP systems is called meta-stars. Its basic idea is to use meta-modeling coupled with navigation tables and with dimension tables: navigation tables support hierarchy instances with different lengths and with non-leaf facts, and allow different roll-up semantics to be explicitly annotated; meta-modeling enables hierarchy heterogeneity and dynamics to be accommodated; dimension tables are easily integrated with standard business hierarchies. After outlining a reference architecture for social business intelligence and describing the meta-star approach, we formalize its querying expressiveness and give a cost model for the main query execution plans. Then, we evaluate meta-stars by presenting experimental results for query performances and disk space.

1. Introduction

The planetary success of social networks and the widespread diffusion of portable devices has enabled simplified and ubiquitous forms of communication and has contributed, during the last decade, to a significant shift in human communication patterns towards the *voluntary sharing of personal information*. Most of us are able to connect to the Internet anywhere, anytime, and continuously send messages to a virtual community centered around blogs, forums, social networks, and the like. This has resulted in the accumulation of enormous amounts of *user-generated content* (UGC), that include geolocation, preferences, opinions, news, etc. This huge wealth of information about people's tastes, thoughts, and actions is obviously raising an increasing interest from decision makers because it can give them a fresh and timely perception of the market mood; besides, often the diffusion of UGC is so widespread to directly influence in a decisive way the phenomena of business and society [1–3].

Some commercial tools are available for analyzing the UGC from a few predefined points of view (e.g., brand reputation and topics correlation) and using some ad hoc KPIs (e.g., topic presence counting and topic sentiment). These tools do not rely on any standard data schema; often they do not even lean on a relational DBMS but rather on in-memory or non-SQL ones. Currently, they are perceived by companies as self-standing applications, so UGC-related analyses are run separately

* Corresponding author.
E-mail addresses: enrico.gallinucci2@unibo.it (E. Gallinucci), matteo.golfarelli@unibo.it (M. Golfarelli), stefano.rizzi@unibo.it (S. Rizzi).

http://dx.doi.org/10.1016/j.is.2015.04.005
0306-4379/© 2015 Elsevier Ltd. All rights reserved.

from those strictly related to business, that are carried out based on corporate data using traditional business intelligence platforms. To give decision makers an unprecedentedly comprehensive picture of the ongoing events and of their motivation, this gap must be bridged [4].

How to extract most information out of the UGC and use it is a hot research theme in different areas, such as information retrieval, text mining, and natural language processing; each community contributes to this common goal by employing different techniques. The perspective we focus on in this paper is that of *social business intelligence* (SBI), that is the discipline of effectively and efficiently combining corporate data with UGC to let decision-makers analyze and improve their business based on the trends and moods perceived from the environment [5]. The data to be combined have very different features: while corporate data are structured, reliable, and accurate, UGC is unstructured or poorly structured, possibly fake, often vague and imprecise; however, both types of data are crucial for an effective decision-making process. As in traditional business intelligence, the goal of SBI is to enable powerful and flexible analyses for users with a limited expertise in databases and ICT; this goal is typically achieved by storing information into a data warehouse, in the form of multidimensional cubes to be accessed through OLAP techniques.

In the context of SBI, the category of UGC that most significantly contributes to the decision-making process in the broadest variety of application domains is the one coming in the form of textual *clips* [2,3]. Clips can either be messages posted on social media (such as Twitter, Facebook, blogs, and forums) or articles taken from on-line newspapers and magazines. Digging information useful for decision-makers out of textual UGC requires first crawling the web to extract the clips related to a *subject area*, then enriching them in order to let as much information as possible emerge from the raw text. The subject area defines the project scope and extent, and can be for instance related to a brand or a specific market. Enrichment activities range from the simple identification of relevant parts (e.g., author, title, and language) if the clip is semi-structured, to the use of either natural language processing or text analysis techniques to interpret each sentence and if possible assign a polarity to it (i.e., *sentiment analysis* or *opinion mining* [6]). Though the issues related to the overall process have been thoroughly investigated in the literature starting from the early 2000s and some commercial tools are available to support all or parts of it, the analysis capabilities of the results delivered to end-users are typically very limited: only static or poorly flexible reports are provided, and historical data are not made available. Besides, in standard architectures the flow of textual UGC is separate from the ETL flows carrying business data, which forces an unnatural dividing line in the decision-making process and dramatically reduces its effectiveness.

1.1. From topics to topic hierarchies

A key role in the analysis of textual UGC is played by *topics*, meant as specific concepts of interest within the subject area [4]. Users are interested in knowing how much people talk about a topic, which words are related to it, if it has a good or bad reputation, etc. Thus, topics are obvious candidates to become a dimension of the cubes for SBI. In this subsection we explain what we mean by topic and topic hierarchy to avoid misunderstandings due to heterogeneous backgrounds of readers.

A topic could be a word having a specific role in the users' business glossary (e.g., a product, a product type, or a brand), or it could be a common word that at some time becomes relevant to the subject area. In SBI projects, a first list of relevant topics and relationships is often manually provided by decision makers and by experts of the subject area, to be then iteratively refined and enriched by analyzing the dynamics of the subject area [7]. In other situations, this task is automated by employing topic discovery algorithms (e.g., [8–10]). When topics are manually defined, their relevance to the subject area is normally higher [11]. Conversely, if topics are automatically discovered, we can expect a wide range of heterogeneous concepts (to be then manually restricted to better focus analyses). For example, COBRA [12] by IBM mixes the two approaches to identify the set of topics that define a company profile at best. Depending on the tool or technique adopted for UGC analysis, each topic is normally coupled with some measures taken either at the clip/sentence level (e.g., number of occurrences of that topic in each clip) or for each single occurrence (e.g., sentiment of each occurrence of that topic in each clip). Such a detailed information is useful, e.g., for early-alerting applications [11] in which users need to timely react to some specific message; however, to effectively summarize the mood raised by a topic, topic measures must be aggregated using clip metadata, e.g., by author, media type, or language, which can be easily done through OLAP-style analyses.

On the other hand, OLAP analyses of single, specific topics are often not sufficient to give users a clear and comprehensive picture of the social mood. Like for any other dimension, users are also very interested in grouping topics together in different ways to carry out more general and effective analyses—which requires the definition of a topic hierarchy that specifies inter-topic roll-up (i.e., grouping) relationships so as to enable aggregations of topic measures at different levels. Hierarchically organized topics have been found to be useful in many contexts such as group profiling at varying granularity [13] and semantic comparison of documents [14]. Many solutions to create such hierarchies have been proposed: in some cases higher hierarchy levels are associated with topics occurring more frequently [15], while in others [16] higher levels are associated to more general topics from an ontological point of view (e.g., the arcs in the hierarchy represent part-of and instance-of relationships). Consistently with the OLAP metaphor, we opt for the second interpretation of hierarchy. Like for topic discovery, a topic hierarchy can be manually defined, or it can be automatically derived from the business glossary or from a mining process. For example, in [16] a methodology is proposed to extract information from big data, convert it into a human-comprehensible format, and build a hierarchical ontological tree based on the extracted metadata.

Once a topic hierarchy is available, it can be used to compute aggregated measures. In particular, the way how an *aggregated sentiment* is computed depends on how the sentiment of each topic occurrence is returned by the sentiment