SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence

Rafael Berlanga, Universitat Jaume I, Castellón de la Plana, Spain
Lisette Garcia-Moya, Universitat Jaume I, Castellón de la Plana, Spain
Victoria Nebot, Universitat Jaume I, Castellón de la Plana, Spain
Maria José Aramburu, Universitat Jaume I, Castellón de la Plana, Spain
Ismael Sanz, Universitat Jaume I, Castellón de la Plana, Spain
Dolores María Llidó, Universitat Jaume I, Castellón de la Plana, Spain

ABSTRACT

The tremendous popularity of web-based social media is attracting the attention of the industry to take profit from the massive availability of sentiment data, which is considered of a high value for Business Intelligence (BI). So far, BI has been mainly concerned with corporate data with little or null attention to the external world. However, for BI analysts, taking into account the Voice of the Customer (VoC) and the Voice of the Market (VoM) is crucial to put in context the results of their analyses. Recent advances in Sentiment Analysis have made possible to effectively extract and summarize sentiment data from these massive social media. As a consequence, VoC and VoM can be now listened from web-based social media (e.g., blogs, reviews forums, social networks, and so on). However, new challenges arise when attempting to integrate traditional corporate data and external sentiment data. This paper deals with these issues and proposes a novel semantic data infrastructure for BI aimed at providing new opportunities for integrating traditional and social BI. This infrastructure follows the principles of the Linked Open Data initiative.

Keywords: Linked Open Data, Opinions, Sentiment Analysis, SLOD-BI Datasets, Social Business Intelligence

1. INTRODUCTION

The massive adoption of web-based social media for the daily activity of e-commerce users, from customers to marketing departments, is attracting more and more the attention of Business Intelligence (BI) companies. So far BI has been confined to corporate data, with little attention to external data. Capturing external data for contextualizing data analysis operations is a time-consuming and complex task that, however, would bring large benefits to current BI environments (Pérez et al., 2008a). The main external contexts for e-commerce ap-
Applications are the Voice of the Customer (VoC) and the Voice of the Market (VoM) forums. The former regards the customer opinions about the products and services offered by a company, and the latter comprises all the information related to the target market that can affect the company business. Listening to the VoM allows setting the strategic direction of a business based on in-depth consumer insights, whereas listening to the VoC helps to identify better ways of targeting and retaining customers. As pointed out by Reidenbach (2009), both perspectives are important to build long-term competitive advantage.

The traditional scenario for performing BI tasks has dramatically changed with the consolidation of the Web 2.0, and the proliferation of opinion feeds, blogs, and social networks. Nowadays, we are able to listen to the VoM and VoC directly from these new social spaces thanks to the burst of automatic methods for performing sentiment analysis over them (Liu, 2012). These methods directly deal with the posted texts to identify global assessments (i.e., reputation) over target items, to detect the subject of the opinion (i.e., aspects) and its orientation (i.e., polarity). From now on, we will consider as social data the collective information produced by customers and consumers as they actively participate in online social activities, and we will refer to all the data elements extracted from social data by means of sentiment analysis tools as sentiment data.

A good number of commercial tools have recently appeared in the market for listening and analyzing social media and product review forums, for example Salesforce Radian6 (http://www.salesforce.com/marketing-cloud), Media Miser (http://www.mediamiser.com), and Synthesio (http://synthesio.com), to mention just a few. Unfortunately, these commercial tools aim to provide customized reports for end-users, and sentiment data on which these reports rely on are not publicly available (indeed this is the key of their business). Consequently, critical aspects such as the quality and reliability of the delivered data cannot be contrasted nor validated by the analysts. This fact contrasts with the high quality that BI requires for corporate data in order to make reliable decisions.

Apart from the sentiment analysis approaches, there is also a great interest on publishing strategic data for BI tasks within the Linked Open Data (LOD) cloud (Heath & Bizer, 2011). The Web 3.0 and LOD are about publishing data identified and linked to each other through a Unique Resource Identifier (URI), and providing data with well-defined semantics to allow users and machines to rightly interpret them. Projects like Schema.org are allowing the massive publication of product offers as micro-data, as well as specific vocabularies for e-commerce applications. Unfortunately, nowadays there is no open data infrastructure that allows users and applications to directly perform analysis tasks over huge amounts of published opinions in the Web.

In this paper we discuss the opportunities and advantages of defining new data infrastructures for performing social BI. As Figure 1 shows, in this social BI infrastructure, VoC and VoM sentiment data must be integrated together with all the external factors that may potentially affect a business (e.g., new legislations, financial news, etc.). We claim that such a data infrastructure must follow the principles of the LOD initiative. As a result, if web-based social data is migrated to the Web 3.0 as linked data in order to be shared, validated and eventually integrated with corporate data, a new global BI scenario for e-commerce applications is enabled. Furthermore, most data and vocabularies used by researchers and companies for performing sentiment analysis could be better exploited if they are shared, contrasted and validated by the community.

The main contributions of this paper are summarised as follows:

- We propose BI analytical patterns to combine corporate with social data;
- We propose a novel semantic data infrastructure to publish both social data and automatically extracted sentiment data. This data infrastructure follows the LOD principles, and therefore it is aimed at...