Process Model for Content Extraction from Weblogs

Andreas Schieber, University of Technology Dresden, Dresden, Germany
Andreas Hilbert, University of Technology Dresden, Dresden, Germany

ABSTRACT

This paper develops and evaluates a BPMN-based process model which identifies and extracts blog content from the web and stores its textual data in a data warehouse for further analyses. Depending on the characteristics of the technologies used to create the weblogs, the process has to perform specific tasks in order to extract blog content correctly. The paper describes three phases: extraction, transformation and loading of data in a repository specifically adapted for blog content extraction. It highlights the objectives in these phases which must be achieved to ensure the correct extraction. The authors integrate the described process in a previously developed framework for blog mining. The authors’ process model closes the conceptual gap in this framework as well as the gap in current research of blog mining process models. Furthermore, it can easily be adapted for other web extraction proposals.

Keywords: BPMN, Data Warehouse, ETL, Process Model, Social BI, Web Content Extraction, Weblog

1. INTRODUCTION

The development of the World Wide Web led to an increasing number of weblogs and their posts in recent years (Baloglu, Wyne, & Bachtetepe, 2010; Chau, Lam, Shiu, Xu, & Jinwei Cao, 2009; WordPress, 2013b; WordPress, 2013a). The authors of these weblogs (known as blogs) also discuss products or services and publish their opinions on how to improve a company’s product. Therefore, more and more companies run their own blog using it as a further communication channel to the customers to receive important information for the design of their products (Lakshmanan & Oberhofer, 2010; Chau et al., 2009; Kaiser, 2009b). Davis and Oberholtzer (2008) also indicate that blogs are an important source for the collection of customer feedback and hence a valuable complement to traditional market research.

The number of existing blogs and posts has now become unmanageable (Chau et al., 2009). As a result the identification and investigation of relevant blogs and content is restricted. Many academic studies deal with the analysis of weblogs. Using the methods of business intelligence (BI) such as data, text and web mining, essential items can be separated from the widespread crowd. Implementing BI’s capabilities in this context, a new research area called Social BI was established which focuses on the analysis of social web data (Dinter & Lorenz, 2012; Zeng, Chen, Lusch, & Li, 2010).

DOI: 10.4018/ijiit.2014040102
For analyzing the information from blog posts by means of (Social) BI, the content of the posts must be collected from the web. Considering traditional BI systems this step is provided by the so called ETL process which extracts, transforms, and loads (ETL) data from different sources into a data warehouse (Akkaoui, Mazón, Vaisman, & Zimányi, 2012; Chaudhuri, Dayal, & Narasayya, 2011). For analyses, like opinion mining itself, it is not necessary for the blog data to be stored in a data warehouse (DWH). However, the steps which have to be executed to receive the data are the same as the steps to store the data in a DWH. As an advantage, a DWH stores historical data, provides fast analyses as well as time series analysis.

ETL processes filling DWHs with blog data are not yet generally described by researchers. Dinter and Lorenz (2012) on the other hand, explicitly mention the research demand for ETL techniques in the context of Social BI because of the specific characteristics of social media and its content. With this in mind and considering the previously developed framework for blog analysis (Schieber & Hilbert, 2009), this paper’s objective is to develop a conceptual model for an ETL process. We describe the three phases, namely, ‘extraction’, ‘transformation’, and ‘loading’ of data in a repository which is specifically adapted for blog content extraction. The focus is on the tasks which must be performed in these phases to ensure the correct extraction.

Following a design-science-based research approach (Hevner, March, Park, & Ram, 2004; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007) we refine our previous work with a detailed conceptual model of the ETL tasks. We derived this concept by analyzing existing academic literature. Firstly, we identified tasks from related work in the research areas of web content extraction and textual data processing. Afterwards we summarized and grouped these tasks into steps, and organized them in a reasonable order as they should be performed within the ETL process. As a result, we visualize our findings in a BPMN-based process model determining the tasks for each phase within the ETL process. Following the proposal by Akkaoui et al. (2012), we use the BPM-notation because the function, which an ETL process fulfills, can be compared with traditional business processes adding value by performing specific tasks. For modeling such processes, BPMN is a standardized and widely used notation. The developed model presents the tasks which have to be completed in order to identify and extract weblog content and to store it in a DWH. This enables further analyses of the collected blog posts.

The paper first discusses related work in our research area. Afterwards we describe some foundations which are relevant for our further work. On the one hand we target basic blog technologies which are available at the majority of weblog implementations (section 3). On the other hand we focus on the repository which stores the blog data and describes possible technologies for its implementation (section 4). Section 5 deals with the functions of an ETL process in order to extract textual data from blogs, transform and adjust it, and finally load it into a DWH. The overall model providing the relationships between the single process models and an implementation of a prototype system are presented in section 6. Finally, we summarize our work, highlight limitations and provide an outlook for future research.

2. BACKGROUND

A lot of research has been already conducted in the area of blog analysis. For example, many authors explain how blog posts can be analyzed to discover trends and opinions published by blog owners or readers. Some authors describe systems which are able to show trends in the blogosphere (Sommer, Schieber, Hilbert, & Heinrich, 2012; Bansal & Koudas, 2007; Glance, Hurst, & Tomokiyo, 2004). Thus, topics are identified, which attract the attention of the blog authors. In contrast, some authors use the textual blog content to extract sentiments and opinions related to products or events to show negative and positive user expressions (Schieber, Hilbert, & Stillich, 2012; Xu et al., 2011; Feng, Wang, Yu, Yang, & Yang, 2009; Kaiser, 2009a).
Sequence Processing with Recurrent Neural Networks
[www.igi-global.com/chapter/sequence-processing-recurrent-neural-networks/10424?camid=4v1a](www.igi-global.com/chapter/sequence-processing-recurrent-neural-networks/10424?camid=4v1a)

Relative Superiority of Key Centrality Measures for Identifying Influencers on Social Media
[www.igi-global.com/article/relative-superiority-of-key-centrality-measures-for-identifying-influencers-on-social-media/123941?camid=4v1a](www.igi-global.com/article/relative-superiority-of-key-centrality-measures-for-identifying-influencers-on-social-media/123941?camid=4v1a)