Chapter 15
Adaptable Services for Novelty Mining

Flora S. Tsai  
Nanyang Technological University, Singapore

Agus T. Kwee  
Nanyang Technological University, Singapore

Wenyin Tang  
Nanyang Technological University, Singapore

Kap Luk Chan  
Nanyang Technological University, Singapore

ABSTRACT

Novelty mining is the process of mining relevant information on a given topic. However, designing adaptable services for real-world novelty mining faces several challenges like real-time processing of incoming documents, computational efficiency, multi-user working environment, diverse system requirements, and integration of domain knowledge from different users. In this paper, the authors bridge the gap between generic data mining methodologies and domain-specific constraints by providing adaptable services for intelligent novelty mining that model user preferences by synthesizing the parameters of novelty scoring, threshold setting, performance monitoring, and contextual information access. The resulting novelty mining system has been tested in a variety of performance situations and user settings. By considering the special issues based on domain knowledge, the authors’ adaptable novelty mining services can be used to support a real-life enterprise.

INTRODUCTION

With the fast growth of technology, the Web is changing from a data-centric Web into Web of semantic data and Web of services (Yee, Tiong, Tsai, & Kanagasabai, 2009). Moreover, the demand for Web services that enable users to run offline standalone applications over the Internet has increased rapidly. The World Wide Web Consortium (W3C) defines a Web service as “a software system designed to support interoperable machine-to-machine interaction over a network” (Hugo & Allan, 2004). As a result, recently more and more software applications are available online and can be accessed on the remote system at the client side. To easily identify those online applications, each of them is assigned with the unique URI (Uniform Resource Indicator), which serves as the address of the individual application or service. Individual services can be built and combined with each other to create other services with more comprehensive functionality at little
Adaptable Services for Novelty Mining

additional cost (Zheng & Bouguettaya, 2009). The use of these Web services has significance in the business domain, where they are used as means of communication or exchanging data between businesses and clients (Kwee & Tsai, 2009).

Another consequence of the rapid growth of technology is the information overload from news articles, scientific papers, blogs (Chen, Tsai, & Chan, 2007), social networks (Tsai, Han, Xu, & Chua, 2009), and mobile content (Tsai et al., 2010). Because of the vast amount of information available, people tend to suffer from information overload because of irrelevant and redundant information in these documents. Thus, novelty mining (NM), or novelty detection, is a solution to this phenomenon. A novelty mining process is a process of retrieving novel yet relevant information, based on a topic given by the user (Ng, Tsai, & Goh, 2007; Ong, Kwee, & Tsai, 2009). Novelty mining can be used to solve many solve many business problems, such as in corporate intelligence (Tsai, Chen, & Chan, 2007) and cyber security (Tsai, 2009; Tsai & Chan, 2007). The pioneer work of novelty mining was proposed by Zhang et al. at the document level (Zhang, Callan, & Minka, 2002). They defined the definition of “novelty” which was the opposite of “redundancy”. Given any set of documents, a document which was less similar to its history documents was regarded as a “novel” document. Although users can retrieve all the novel documents, each document still needs to be read to find the novel sentences within these documents (Tsai & Chan, 2010). Therefore, to serve users better, later studies of novelty mining were performed at the sentence level (Allan, Wade, & Bolivar, 2003; Kwee, Tsai, & Tang, 2009; Tang & Tsai, 2010; Zhang, Xu, Bai, Wang, & Cheng, 2004; Zhang & Tsai, 2009b). To the best of our knowledge, no previous work has been reported in designing and developing adaptable services for novelty mining system in the business enterprise. For enterprise users, this service-oriented novelty mining system conveniently helps them to retrieve new information about certain events of interest. They do not need to read through all documents or passages in order to find the novel information. Creating Web services for the novelty mining components allows for the rapid deployment and availability of these services for these diverse set of users, which can balance technical significance and business concerns in business processes and enterprise systems.

This paper addressed the problem of intelligent novelty mining, which is a big challenge in the data mining and business community. Intelligent novelty mining solves the domain-specific problem of mining novel information from text data with specific regard to the user context. Traditional novelty mining focuses on optimizing the retrieval of novel information with a predefined evaluation measure. This tends not be useful in realistic situations where a fixed unit of measure is unable to adapt to the performance requirements of different users. Intelligent novelty mining, on the other hand, aims to balance the technical significance and business concerns to create techniques that are useful in real-world scenarios. We describe the adaptable services used in an intelligent novelty mining system that is used to detect and retrieve any new information of an incoming stream of text. More specifically, the contributions are to propose services for each aspect of novelty mining, which includes novelty scoring service for mixed metrics, adaptable threshold setting service, novelty feedback service for contextual information, and document-to-sentence service for improved document-level detection.

LITERATURE REVIEW

Document-Level Novelty Mining

Zhang et al. (2002) performed document-level novelty mining based on the distance between the new document and the previously delivered documents (in history), i.e., document-document distance. This method simulates the way users