Chapter 13
Personalizing News Services Using Semantic Web Technologies

Flavius Frasincar  
Erasmus University Rotterdam, The Netherlands

Jethro Borsje  
Erasmus University Rotterdam, The Netherlands

Frederik Hogenboom  
Erasmus University Rotterdam, The Netherlands

ABSTRACT

This chapter describes Hermes, a framework for building personalized news services using Semantic Web technologies. The Hermes framework consists of four phases: classification, which categorizes news items with respect to a domain ontology; knowledge base updating, which keeps the knowledge base up-to-date based on the news information; news querying, which allows the user to search the news with concepts of interest; and results presentation, which shows the news results of the search process. Hermes is supported by a framework implementation, the Hermes News Portal, a tool that enables users to have a personalized access to news items. The Hermes framework and its associated implementation aim at advancing the state-of-the-art of semantic approaches for personalized news services by employing Semantic Web standards, exploiting and keeping up-to-date domain information, using advanced natural language processing techniques (e.g., ontology-based gazetteering, word sense disambiguation, etc.), and supporting time-based queries for expressing the desired news items.

INTRODUCTION

The simplicity, availability, reachability, and reduced exploitation costs have made the Web one of the most common platforms for information publishing and dissemination. This is particularly true for news agencies that use Web technologies to present emerging news regarding different types of events as for example business, cultural, sport, and weather events. Most of this information is published as unstructured text that is made available to a general audience by means of Web pages.
The heterogeneity of the Web audience and the diversity of the published information asks for more refined ways of delivering information that would enable users to access news items that interest them. For this purpose the Really Simple Syndication (RSS) (Winer, 2003) standard was developed that publishes information in a semi-structured format that supports machine processing. This format is based on metadata that (1) associates news items with channels (feeds) that have properties like categories (e.g., business, sport, politics, etc.), title, publication date, etc., and (2) describes news items by means of their properties as categories (e.g., online business, business system, Internet marketing, etc.), release time, title, abstract, link to the original published information, etc.

Most of the annotations supported by the RSS feeds are coarse-grained providing general news information. Fine-grained information, as for example the financial events depicted in news, is at the moment not available. Also, the current annotations are only partially processable by machines as the tags do not have formal semantics associated and hence have different interpretations. Being able to understand the semantic content of a news item would enable a fine-grained categorization of this information, thus better supporting the users (casual users, media analysts, stock brokers, etc.) information needs.

In order to make the Web data not only machine readable but also machine understandable the World Wide Web Consortium proposes the Semantic Web (Berners-Lee, Hendler, & Laszil, 2001), a set of technologies that allow for self-describing content. On the Semantic Web, metadata is defined using semantic information usually captured in ontologies. Some of the most popular formats to describe ontologies on the Semantic Web are RDF(S) (Klyne & Carroll, 2004) (Brickley & Guha, 2004) and OWL (Bechhofer et al., 2004).

A special class of users who make daily use of (emerging) news is that of stock brokers. Because news messages may have a strong impact on stock prices, stock brokers need to monitor these messages carefully. Due to the large amounts of news information published on a daily basis, the manual task of retrieving the most interesting news items with respect to a given portfolio is a challenging one. Existing approaches such as Google Finance or Yahoo! Finance are developed to meet these personalization needs by supporting automatic news filtering on the Web.

Current approaches to news filtering are able to retrieve only the news that explicitly mention the companies involved, failing to deliver indirect information which is also deemed relevant for the considered portfolio. For example, for a portfolio based on Google shares, such systems fail to deliver news items related to competitors of Google, such as Yahoo! or Microsoft, which might have an indirect influence on the share price of Google. Exploiting the semantic contextual information related to companies such as its competitors, CEO’s, alliances, products, etc., enables a more comprehensive overview of relevant news with respect to a certain portfolio.

Existing news filtering systems are not able to cope with delivering news items satisfying temporal constraints. The time aspect is of utmost importance when, for example, one considers the fact that news items usually have an immediate impact on stock prices, or when one desires to do a historical analysis of past news and stock price evolutions. Being able to exploit the timestamps associated to news items enables retrieving only news that obey user-determined time-related constraints.

Another limitation of existing news personalization services is that they fail to consider the dynamicity of the current world. In an economic context companies are created, disappear, or are bought by other companies, CEO’s change positions or sometimes even companies, companies develop new products while other products become obsolete, etc. For a proper representation of the domain it is important that we cope with