ABSTRACT

This chapter presents a methodology for personalized knowledge discovery from text. Traditionally, problems with text mining are numerous rules derived and many already known to the user. Our proposed algorithm derives user’s background knowledge from a set of documents provided by the user, and exploits such knowledge in the process of knowledge discovery from text. Keywords are extracted from background documents and clustered into a concept hierarchy that captures the semantic usage of keywords and their relationships in the background documents. Target documents are retrieved by selecting documents that are relevant to the user’s background. Association rules are discovered among noun phrases extracted from target documents. Novelty of an association rule is defined as the semantic distance between the antecedent and the consequent of a rule in the background knowledge. The experiment shows that our novelty measure performs better than support and confidence in identifying novel knowledge.

INTRODUCTION AND BACKGROUND

The goal of Text Mining is to find interesting and non-trivial knowledge from unstructured documents. It can be viewed as the extension of Data Mining to textual data. Data mining tools are capable of discovering unknown knowledge from huge amount of data, but most of the time the number of discovered patterns is too large for a user to find interesting patterns quickly and easily from the result. In a study
conducted by Stanford University, the association rule mining algorithm generated over 20,000 rules from a subset of the census data containing about 30,000 records. Most of the rules are not useful, and those “that came out at the top, are things that were obvious” (Brin, Motwani, Ullman, Tsur, 1997). In Text Mining, the problem becomes even more critical because of the large number of documents available and the high dimensionality of textual data. Identifying interesting rules for a particular user has become a major issue for text mining research.

In Data Mining field, both objective and subjective measures have been proposed to evaluate the interestingness of discovered patterns (Liu, Hsu, & Ma, 2001; Padmanabhan & Tuzhilin, 1999; Piatetsky-Shapiro, & Matheus, 1994; Silberschatz, & Tuzhilin, 1995). However, objective measures alone are insufficient, because they rely only on the characteristics (surface features) of the patterns and the underlying data collection without considering users’ knowledge and interests. One can generate a large number of rules that are interesting “objectively” but of little interest to the user (Klemettinen, Mannila, Ronkainen, Toivonen, & Verkamo, 1999). Subjective measures are not easy to implement in practice because of the difficulty to obtain users’ subjective opinions. Subjective measures, such as unexpectedness (a pattern is interesting if it is “surprising” to the user) and actionability (a pattern is interesting if the user can act on it to his/her benefit) (Silberschatz, & Tuzhilin, 1995), assess the interestingness of patterns from the users’ perspective, but explicit expressions of users’ interests (expectations/unexpectations) are required in order to perform the comparison. In practice it is difficult or even nearly impossible for users to do so, especially before the discovered patterns are presented to the user.

In this chapter, we propose a text mining technique that discovers personalized knowledge from large document collections. The system derives a user’s background knowledge implicitly from a set of documents that are already known to the user (aka background documents). The user’s background knowledge is represented as a key word space containing a concept hierarchy developed from key words extracted from background documents. The background knowledge is then used to retrieve documents that are relevant to the user’s background (aka target documents) from a large corpus. The knowledge to be discovered is in the form of association rules. Noun phrases are extracted from target documents, and association rules are mined among noun phrases. Interesting rules are identified by comparing the discovered association rules to the user’s background knowledge. A novelty measure, defined as the semantic distance between the antecedent and the consequent of a rule in the background key word space, is used to predict the novelty of an association rule. Because the target document formulation and the novelty calculation are determined by the background documents, the discovered association rules are customized for each user. The experiment result shows that the novelty measure is highly correlated with the subjective ratings of the novelty of association rules. There is a high correlation between the automatic novelty measure and the subjective usefulness ratings as well. Overall the novelty measure performs significantly better than the support and the confidence measure in terms of identifying novel and useful knowledge.

The remainder of the chapter describes the proposed methodology for deriving users’ background knowledge, retrieving the target documents, extracting features from target documents, discovering novel association rules, and the user experiment.
Related Content

Working from Claims Data
www.igi-global.com/chapter/working-claims-data/36640?camid=4v1a

Automated Integration of Heterogeneous Data Warehouse Schemas
www.igi-global.com/article/automated-integration-heterogeneous-data-warehouse/1815?camid=4v1a

Agricultural Data Mining in the 21st Century
www.igi-global.com/chapter/agricultural-data-mining-21st-century/29153?camid=4v1a

A Cached-Based Stream-Relation Join Operator for Semi-Stream Data Processing
www.igi-global.com/article/a-cached-based-stream-relation-join-operator-for-semi-stream-data-processing/168484?camid=4v1a