Chapter 36
Visualization of High-Level Associations from Twitter Data

Luca Cagliero
Politecnico di Torino, Italy

Naeem A. Mahoto
Politecnico di Torino, Italy

ABSTRACT

The Data Mining and Knowledge Discovery (KDD) process focuses on extracting useful information from large datasets. To support analysts in making decisions, a relevant research effort has been devoted to visualizing the extracted data mining models effectively. A particular attention has been paid to the discovery of strong association rules from textual data coming from social networks, which represent potentially relevant correlations among document terms. However, state-of-the-art rule visualization tools do not allow experts to visualize data correlations at different abstraction levels. Hence, the effectiveness of the proposed approaches is limited, especially when dealing with fairly sparse data. This chapter presents Twitter Generalized Rule Visualizer (TGRV), a novel text mining and visualization tool. It aims at supporting analysts in looking into the results of the generalized association rule mining process from textual data coming from Twitter supplied with WordNet taxonomies. Taxonomies are used for aggregating document terms into higher-level concepts. Generalized rules represent high-level associations among document terms. By exploiting taxonomy-based models, experts may look into the discovered data correlations from different perspectives and figure out interesting knowledge. Changing the perspective from which data correlations are visualized is shown to improve the readability and the usability of the generated rule-based model. The experimental results show the applicability and the usefulness of the proposed visualization tool on real textual data coming from Twitter. The visualized data correlations are shown to be valuable for advanced analysis, such as topic trend and user behavior analysis.

INTRODUCTION

Data Mining and Knowledge Discovery (KDD) focuses on extracting useful information from large datasets (Tan & al., 2005). Descriptive data mining techniques (e.g., clustering, association rule mining) entail discovering interesting and hidden patterns from the analyzed data. In the last several years a significant research effort has been devoted to applying data mining techniques to textual data published on social networks. In particular, the analysis of the textual User-

DOI: 10.4018/978-1-4666-9562-7.ch036

Generated Content (UGC) published on Twitter (http://twitter.com) has achieved promising results in the context of user behavior profiling (Li et al., 2008; Mathioudakis & Koudas., 2010) and topic trend discovery (Cheong & Lee., 2009; Cagliero & Fiori, In press).

Association rule mining (Agrawal & al., 1993) is a widely exploratory data mining technique that allows discovering valuable correlations among data. An association rule is an implication A → B, where A and B are sets of items occurring in the source data. In the context of textual data analysis, a rule represents an implication between a couple of term sets occurring in the analyzed document. To make the rule mining process computationally tractable, a minimum support threshold is commonly enforced to select only the associations among terms that occur frequently in the analyzed data. As a drawback, traditional rule mining algorithms (e.g., Apriori (Agrawal & Srikant, 1994), FP-Growth (Han et al., 2000)) are sometimes ineffective in mining valuable knowledge, because of the excessive level of detail of the mined information. For instance, when coping with real-world textual data, most of the associations among terms occur rarely in the analyzed data and, thus, may be discarded by enforcing a minimum support threshold. To overcome this issue, Agrawal & Srikant (1995) proposed to discover generalized association rules. Generalized rules are rules that may also contain high level (generalized) terms. By exploiting a taxonomy (i.e., a set of is-a hierarchies) built over the analyzed textual documents terms are aggregated into higher level concepts, which are more likely to be frequent in the analyzed data. Hence, generalized rules represent underlying term correlations at different abstraction levels. Generalized rule mining from textual data has already been addressed in different application contexts, among which social data analysis (Cagliero & Fiori, In Press) and biomedical literature analysis (Berardi et al., 2005).

To support analysts in the knowledge discovery process a parallel relevant research effort has been devoted to proposing visual tools adapted to several well-known KDD tasks. In the context of association rule mining, the proposed systems are commonly focused on either visualizing the mining results effectively to ease the expert validation task (Leung et al. 2008; Wong et al., 1999; Meng, 2010) or allowing experts to drive the data mining process (Fayyad et al., 2001; Li et al. 2011). However, to the best of our knowledge, the problem of visualizing generalized rules mined from textual data coming from social networks has never been investigated so far.

This Chapter presents a novel visualization tool, called *Twitter Generalized Rule Visualizer* (TGRV), which allows experts to explore the result of the generalized rule mining process from textual data coming from Twitter effectively. Twitter textual messages published by Web users and ranging over the same topic are retrieved through the Twitter Application Programming Interfaces (APIs) and integrated in common repositories. Next, frequent generalized association rules are discovered from the generated datasets by exploiting a WordNet taxonomy built on the analyzed data. Finally, a graph-based rule visualization model, namely the *Generalized Rule Graph*, is generated to allow experts to explore the mined rules from different perspectives. Generalized Rule Graph nodes represent term sets of arbitrary size, while the oriented edges represent strong associations between node couples. Analyzing the Generalized Rule Graph from different perspectives allows analysts to avoid exploring the whole set of frequent rules. For instance, changing the abstraction level at which rules are analyzed allows experts to overcome the limitations of traditional rule visualizers, which often provide unsatisfactory results when coping with fairly sparse data (Meng, 2010).

The applicability of the proposed system has been evaluated on real textual data coming from