Expert Systems with Applications 41 (2014) 4950–4958

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Sentimental causal rule discovery from Twitter

Rahim Dehkharghani *, Hanefi Mercan, Arsalan Javeed, Yucel Saygin

Computer Science and Engineering Program, Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Turkey

ARTICLE INFO

Keywords:
Sentiment analysis
Data mining
Machine learning
Causal rules
Sentimental causal rules
Twitter

ABSTRACT

Social media, especially Twitter is now one of the most popular platforms where people can freely express their opinion. However, it is difficult to extract important summary information from many millions of tweets sent every hour. In this work we propose a new concept, *sentimental causal rules*, and techniques for extracting sentimental causal rules from textual data sources such as Twitter which combine sentiment analysis and causal rule discovery. Sentiment analysis refers to the task of extracting public sentiment from textual data. The value in sentiment analysis lies in its ability to reflect popularly voiced perceptions that are stated in natural language. Causal rules on the other hand indicate associations between different concepts in a context where one (or several concepts) cause(s) the other(s). We believe that sentimental causal rules are an effective summarization mechanism that combine causal relations among different aspects extracted from textual data as well as the sentiment embedded in these causal relationships. In order to show the effectiveness of sentimental causal rules, we have conducted experiments on Twitter data collected on the Kurdish political issue in Turkey which has been an ongoing heated public debate for many years. Our experiments on Twitter data show that sentimental causal rule discovery is an effective method to summarize information about important aspects of an issue in Twitter which may further be used by politicians for better policy making.

1. Introduction

Social media platforms offer individuals the opportunity to articulate opinions on various topics ranging from consumer products and services to socio-political issues. These opinions are quite useful in various areas such as marketing for managers or policy making for government agencies.

Sentiment analysis aims to extract opinions towards a topic generally from textual data sources. On the other hand causal rules have been proposed in the literature as an information extraction method in the form of causalities among items in a database. In this paper we combine these two notions, i.e. sentiment analysis and causal rules into "sentimental causal rules" where causal rules based on various aspects extracted from textual data sources are associated with sentiment. We also propose sentimental causal rule discovery techniques from textual data sources.

Sentimental causal rules are an efficient information extraction and mechanism summarizing a large set of textual document into a small subset of rules which can be used for decision making. The problem tackled in this paper is building an information extraction and summarization system which motivated us to propose the current methodology. Our main goal is summarizing a large collection of tweets by several (typically less than 100) sentimental causal rules which can easily be studied by human beings. Such a system can be quite useful for companies and policy makers, for example governments maybe interested in knowing what people think about the political issues.

Twitter is a popular microblogging and social networking website with a registered user base of around 650 millions as of 2013. which allows its users to send text messages of at most 140 characters (*tweets*). Twitter users tweet about everyday subject of life and especially in recent years, for launching political campaigns. Although we applied the mentioned techniques on Twitter, it can be used for other types of data sources.

The rationale for the choice of Twitter as a source of dataset collection lies in its ability to provide a huge number of tweets for the case study of this paper.

In order to demonstrate the effectiveness of sentimental causal rules, we have chosen the Kurdish issue in Turkey as the main topic of Tweets and extracted sentimental causal rules from Twitter on the Kurdish issue in Turkey.

We propose a four-step methodology for sentimental causal rule discovery:

* Corresponding author. Tel.: +90 5342327919.
E-mail addresses: rdehkharghani@sabanciuniv.edu (R. Dehkharghani), hanefimercan@sabanciuniv.edu (H. Mercan), ajaveed@sabanciuniv.edu (A. Javeed), ysaygin@sabanciuniv.edu (Y. Saygin).

http://dx.doi.org/10.1016/j.eswa.2014.02.024
0957-4174/

- The first step is extracting aspect keywords from tweets which will be basis of sentimental causal rules.
- The second step is extracting causal rules among the extracted aspect keywords which frequently appear in tweets.
- The third step is to identify the polarity of tweets as positive, negative, or neutral (objective).
- Finally the forth step assigns polarity values to causal rules based on the aspect keywords in those rules and the polarity of the tweets which support those rules.

Tweets contain aspect keywords related to a topic where these aspect keywords may have polarity which indicates the attitude of the user towards the aspect. Because of the short length of tweets, we assume that the overall polarity of a tweet indicates the polarity of the aspect keywords appeared in it; in other words, the conveyed message by a tweet may include a few aspect keywords, and the message polarity covers the polarity of included aspects.

Our overall methodology for sentimental causal rule discovery is presented in Fig. 1. Two branches in the flowchart indicate two different information extraction tasks – sentiment analysis and causal rule extraction – and the last step in the bottom is combining those tasks to provide a more efficient method.

Although the majority of tweets about the Kurdish political issue in Turkey is in Turkish, in this paper, we have focused only on English tweets to get an idea about international opinion on the matter.

We first extracted aspect keywords from all tweets which were later used in causal rule extraction from Twitter. Our sentiment analysis system labeled tweets as positive, negative, or neutral. After completing the aspect keyword, rule extraction, and sentiment analysis on tweets, we assigned the aspect keyword and rule polarities.

Our main contribution to state of the art is introducing the notion of sentimental causal rules together with a methodology to extract sentimental causal rules from textual data sources. With only pure sentiment analysis on tweet level, the results would be the percentage of positive, negative, or neutral tweets, which itself does not tell much. Also pure causal rule extraction only gives causality relations between different aspects included in a dataset. However, with sentimental causal rules we were able to extract more useful information and see why and in which aspects (and concepts) hold positive or negative opinions by Twitter users. Suggested approach provides summarized information tagged by polarity values. For example "how much an aspect such as *Syria* or a concept included in a causal rule such as *student*, *Kurds* → *attack* has gained positive or negative sentiment by users on Twitter" is the summary of thousands of tweets.

2. Background and preliminaries

In this section, basic definitions and concepts are revised before introducing sentimental causal rules.

Association rules are proposed by Agrawal, Imieliński, and Swami (1993) which are the basis of causal rules. Below a special case of association rules (2-to-1 association rules) are defined based on the association rule definition of Agrawal et al. (1993).

Definition 1. Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of binary attributes, called items and $R = \{\tau_1, \tau_2, \ldots, \tau_n\}$ be a set of transactions where each transaction τ_i contains a set of items $\tau_i = \{i_1, i_2, \ldots, i_k\}$. A 2-to-1 association rule is an implication in the form $i_i, i_j \rightarrow i_k$, where $i_i, i_j,$ and $i_k \in I$ and $i_i \neq i_j \neq i_k$ with support (supp) and confidence (conf) values defined in (1) and (2).

The *significant association rules* (*ARs*) are those with support (*supp*) and confidence (*conf*) values greater than predefined support and confidence thresholds:

$$supp(i_i, i_j \rightarrow i_k) = \frac{|\omega_{ij}|}{|T|} \tag{1}$$

$$conf(i_i, i_j \rightarrow i_k) = \frac{|\omega_{ijk}|}{|\omega_{ij}|} \tag{2}$$

where T is the set of all tweets used in our study, ω_{ij} is a subset of transactions including i_i, and i_j and ω_{ijk} is a subset of transactions including i_i, i_j, and i_k. In the current study, instead of mere associations, we tackle the broader problem of determining causal relations between variables as they infer each other.

In many previous works on association rule mining, only pure association rules as the dependency between variables have been investigated. Those works, however, did not investigate the causalities from extracted rules, such as existence of i_i *and* i_j cause to exist i_k ($i_i, i_j \rightarrow i_k$). Since this causality information can be quite useful in our case (information extraction from political comments), we decided to use this statistical technique.

In our case, we defined a causal rule as two items are implying one ($i_i, i_j \rightarrow i_k$) because of limitations in CCU rule (more detail in Section 4.2.2).

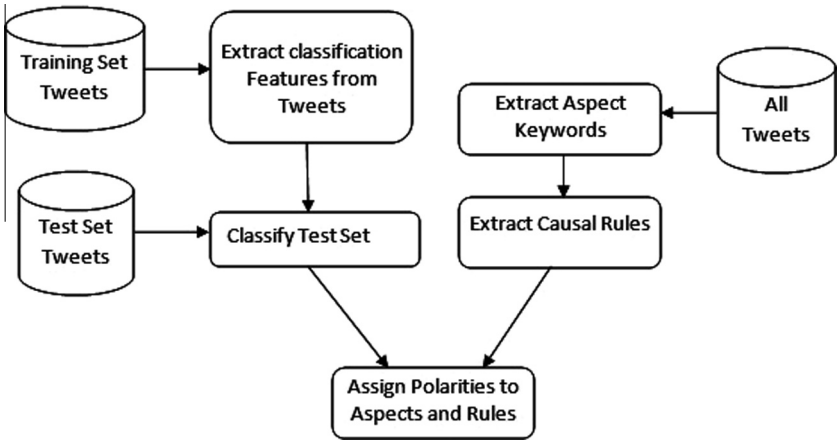

Fig. 1. Overall flowchart of our system for sentimental causal rule discovery.