Expert Systems with Applications 38 (2011) 12094–12100

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A random walk algorithm for automatic construction of domain-oriented sentiment lexicon

Songbo Tan *, Qiong Wu

Key Laboratory of Network, Institute of Computing Technology, Chinese Academy of Sciences, China

ARTICLE INFO

Keywords:
Sentiment analysis
Opinion mining
Information retrieval
Data mining

ABSTRACT

In recent years, many studies have been conducted to deal with automatic construction of domain-oriented sentiment lexicon. However, most of the attempts rely on only the relationship between sentiment words, failing to uncover the mutual relationship between the words and the documents, as well as ignoring the useful knowledge of some existed domains (or "old domain"). This paper proposes a random walk algorithm to construct domain-oriented sentiment lexicon by simultaneously utilizing sentiment words and documents from both old domain and target domain (or "new domain"). The approach simulates a random walk on the graphs that reflect four kinds of relationships (the relationship between words, the relationship from words to documents, the relationship between documents, the relationship from documents to words) between documents and words. Experimental results indicate that the proposed algorithm could dramatically improve the performance of automatic construction of domain-oriented sentiment lexicon.

1. Introduction

Sentiment classification (Tan, Cheng, Wang, & Xu, 2009; Tan, Wu, Tang, & Cheng, 2007; Tan & Zhang, 2008; Tang, Tan, & Cheng, 2009; Wu et al., 2009) is the process of identifying the opinion (e.g. negative or positive) of a given document. With the increasing reviewing pages and blogs etc., this field has received considerable attention from researchers in recent years and has many important applications, such as determining critics' opinions about a given product and summarization (e.g. Ku, Liang, & Chen, 2006).

To date, a variety of methods have been developed for sentiment classification. One of the commonly-used methods is to build a sentiment lexicon. However, it is an impossible task to build a general sentiment lexicon that could perform well in every domain. This is because sentiment expression often behaves with strong domain-specific nature. For example, "portable" and "delicate" are used to express positive sentiment in electronics reviews; while these words are hardly used to convey the same sentiment in hotel reviews. Consequently, the domain-oriented sentiment lexicon is considered as the most valuable resource for sentiment classification tasks.

In recent years, two kinds of approaches have been proposed to deal with this problem: one is based on a thesaurus (Esuli & Sebastiani, 2005; Hu & Liu, 2004; Kamps, Marx, Mokken, & Rijke, 2004; Kim & Hovy, 2004); the other one is based on co-occurrence in a corpus (Du, Tan, Cheng, & Yun, 2010; Gamon & Aue, 2005; Hatzivassiloglous & McKeown, 1997; Kanayama & Nasukawa, 2006; Popescu & Etzioni, 2005; Turney, 2002). However, these kinds of approaches rely on only the labeled seed words to construct a domain-oriented sentiment lexicon, failing to uncover the mutual relationship between the words and the documents; besides, these kinds of approaches ignore the labeled lexicon and labeled corpus in one existed domain (or "old domain").

In fact, the polarity of a sentiment word can be determined by the related documents as well as by the related words, and this rule also holds when determining the polarity of a document. This rule is based on the following intuitive observations:

(1) A word strongly linked with other positive (negative) words could be considered as positive (negative); in the same way, a document strongly linked with other positive (negative) documents could be considered as positive (negative).
(2) A word appearing in many positive (negative) documents could be considered as positive (negative); similarly, a document containing many positive (negative) words could be considered as positive (negative).

Furthermore, labeled data in different domains are very imbalanced. In some traditional domains or domains of concern, many labeled data are freely available on the Web, but in other domains, labeled data are scarce and it involves much human labor to manually label reliable data. As a result, the ideal scheme is to utilize

* Corresponding author. Address: Key Laboratory of Network, P.O. Box 2704, Beijing, 100190, PR China. Tel.: +86 10 62600928; fax: +86 10 62600905.
E-mail address: tansongbo@software.ict.ac.cn (S. Tan).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.105

labeled data in old domain to assist the construction of the domain-oriented lexicon for new domain.

Inspired by these, we aim to take into account all the four kinds of relationships among words and documents (i.e. the relationship between words, the relationship from words to documents, the relationship between documents, and the relationship from documents to words), from old domain as well as from new domain, when constructing a domain-oriented sentiment lexicon.

In this work, we propose an approach based on random walk to implement the above idea. The proposed approach makes full use of all the relationships among words and documents from both old domain and new domain. The detailed procedure can be divided into four steps: Firstly, four graphs are built to reflect the above relationships respectively. Then, we assign a score for every word to denote its extent to "negative" or "positive". Thirdly, we iteratively calculate the score by simulating a random walk on the graphs. Lastly, the final score is achieved when the algorithm converges, so we can determine the semantic orientation of new-domain words based on these scores.

The rest of this paper is organized as follows. Section 2 surveys related work. In Section 3, we introduce the proposed approach in detail. And then Section 4 presents and discusses the experimental results. Lastly we conclude this paper and discuss future work in Section 5.

2. Related work

So far, some studies have been conducted to deal with the lexicon construction problem. Most of these utilize some paradigm words and word similarity to construct lexicon. According to the manner of obtaining word similarity, these methods could roughly be classified into two categories: the first kind of approaches is based on the thesaurus; the second kind is based on the corpus.

2.1. Thesaurus based approach

Thesaurus based approach utilizes synonyms or glosses of a thesaurus to determine polarities of words.

Kim and Hovy (2004) used the synonym and antonym lists obtained from Wordnet to compute the probability of a word given a sentiment class (i.e. positive or negative). Kamps et al. (2004) made a hypothesis that synonyms have the same polarity. And they linked synonyms provided by a thesaurus to build a lexical network, so the word polarity can be determined by the distance from seed words ("good" and "bad") in the network. Hu and Liu (2004) extended the method in Kamps et al. (2004), and they used not only synonyms but also antonyms to build lexical network. Esuli and Sebastiani (2005) determined the orientation of subjective terms based on the quantitative analysis of the glosses of subjective terms, i.e. the definitions that these terms were given in online dictionaries.

2.2. Corpus based approach

Corpus based approach is based on an idea that sentiment words conveying the same polarity co-occur with each other in corpus. Many studies have been done on this field.

Hatzivassiloglous and McKeown (1997) constructed a lexical network from intra-sentential co-occurrence, and identified the positive or negative semantic orientation of the conjoined adjectives. Turney (2002) determined the semantic orientation of a phrase by the mutual information between the given phrase and the word "excellent" minus the mutual information between the given phrase and the word "poor". The mutual information was measured by the number of hits returned by a search engine. Gamon and Aue (2005) extended Turney's approach, and they added one assumption that sentiment words of opposite orientation tended not to co-occur at the sentence level. Popescu and Etzioni (2005) used a relaxation labeling method to find the semantic orientation of words in the context of given product features and sentences. They iteratively assigned semantic orientation labels to words by using various features including intra-sentential co-occurrence and synonyms of a thesaurus. Kanayama and Nasukawa (2006) used both inter-sentential and intra-sentential co-occurrence to get polarities of words and phrases.

However, most of the existed studies rely on only the relationship between words either in a thesaurus (Esuli & Sebastiani, 2005; Hu & Liu, 2004; Kamps et al., 2004; Kim & Hovy, 2004) or in a corpus (Gamon & Aue, 2005; Hatzivassiloglous & McKeown, 1997; Kanayama & Nasukawa, 2006; Popescu & Etzioni, 2005; Turney, 2002), while ignoring other relationships between words and documents (e.g., the relationship from words to documents, and the relationship from documents to words, the relationship between documents and documents), as well as the existed old-domain knowledge.

In order to uncover all these knowledge, in this paper, we design a novel algorithm to construct a domain sentiment lexicon. This algorithm is based on random walk model by taking into account all kinds of relationships among documents and words, from both old domain and new domain.

3. Proposed methods

3.1. Overview

A random walk model assumes that, from one period to the next, the original time series merely takes a random "step" away from its last recorded position (http://www.duke.edu/~rnau/411rand.htm). It has been applied to computer science, physics and a number of other fields.

In the field of information retrieval, given a graph whose edges are weighted by the probability of traversing from one node to another, one critical problem is how to rank the nodes. We can simulate a random walk on the graph. When a walker traverses from one node to another, he visits some notes more often than the others. So we can rank the nodes according to the probabilities that the walker will visit the nodes after the random walk. This idea has been successfully extended in many studies (e.g. PageRank (Brin, Page, Motwami, & Winograd, 1999), LexRank (Erkan & Radev, 2004), CollabRank (Wan & Xiao, 2008)). Our approach is also inspired by this idea.

We can get the following thoughts based on the ideas of PageRank and HITS (Kleinberg, 1998):

(1) If a word is strongly linked with other positive (negative) words, it tends to be positive (negative); and if a document is strongly linked with other positive (negative) documents, it tends to be positive (negative).
(2) If a word appears in many positive (negative) documents, it tends to be positive (negative); and if a document contains many positive (negative) words, it tends to be positive (negative).

Given the data points of words and documents, there are four kinds of relationships in our problem:

- WW-Relationship: It denotes the relationship between words, usually computed by knowledge-based approach or corpus-based approach.
- WD-Relationship: It denotes the relationship from words to documents, usually computed by the relative importance of a document to a word.