Knowledge-Based Systems 76 (2015) 166–175

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Implicit feature identification in Chinese reviews using explicit topic mining model

Hua Xu *, Fan Zhang, Wei Wang

State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 13 November 2014
Received in revised form 6 December 2014
Accepted 10 December 2014
Available online 24 December 2014

Keywords:
Opinion mining
Implicit feature
Topic model
Support vector machine
Product review

ABSTRACT

The essential work of feature-specific opinion mining is centered on the product features. Previous related research work has often taken into account explicit features but ignored implicit features, However, implicit feature identification, which can help us better understand the reviews, is an essential aspect of feature-specific opinion mining. This paper is mainly centered on implicit feature identification in Chinese product reviews. We think that based on the explicit synonymous feature group and the sentences which contain explicit features, several Support Vector Machine (SVM) classifiers can be established to classify the non-explicit sentences. Nevertheless, instead of simply using traditional feature selection methods, we believe an explicit topic model in which each topic is pre-defined could perform better. In this paper, we first extend a popular topic modeling method, called Latent Dirichlet Allocation (LDA), to construct an explicit topic model. Then some types of prior knowledge, such as: must-links, cannot-links and relevance-based prior knowledge, are extracted and incorporated into the explicit topic model automatically. Experiments show that the explicit topic model, which incorporates pre-existing knowledge, outperforms traditional feature selection methods and other existing methods by a large margin and the identification task can be completed better.

1. Introduction

Along with the rapid development of e-commerce, an increasing number of people nowadays prefer to shop on-line. Meanwhile, after purchasing certain products, they have become used to making comments, which can be read by references for other customers. However, as the number of comments gradually increases, the amount of comments for a product may exceed one thousand or even more, which may make potential buyers reluctant to read them all. Furthermore,such a mass of information makes it difficult for the product manufacturers to obtain the practical viewpoints about the product from users. As a consequence, both manufacturers and the customers have difficulty in making appropriate decisions. Therefore, opinion mining, as an emerging technology, is applied to mine the opinions and generate a summary of the products from the immense magnitude of the reviews. There is a significant body of work applying such technology to an wide variety of tasks [17,22,16].

Ding et al. [10] elaborately define feature-specific opinion mining. As mentioned in their paper, the definition of an object is an entity which has been commented on by people. Any product, person, event, organization, or topic, which is associated with a group of attributes and a hierarchy or taxonomy of components, can defined as an object. Meanwhile, each component possesses its own group of subcomponents and attributes. Both components and attributes can be represented by features which are subjects of reviews. For instance, two features are mentioned in the following cell phone review.

Example 1. *Zhe4 Kuan3 Shou3 Ji1 Wai4 Guan1 Hen3 Shi2 Shang4, Ye3 Hen3 Pian2 Yi4.*

(These cell phone is fashionable in appearance, and it is also very cheap.)

A feature is defined as an *explicit feature* when it directly appears in a review. In contrast, when a feature is only implied by other indicators, it is defined as an *implicit feature*. As shown in Example 1, "*appearance*" is an explicit feature, while "*price*", which is implied by "*cheap*", is an implicit feature. Therefore, the

* Corresponding author. Tel.: +86 10 62796450.
E-mail address: xuhua@tsinghua.edu.cn (H. Xu).

http://dx.doi.org/10.1016/j.knosys.2014.12.012
0950-7051/

first sentence, containing at least one explicit feature, is defined as an *explicit sentence*. The second sentence, only containing implicit features, is defined as an *implicit sentence*.

More importantly, daily experiences and statistical data both reveal that a fair amount of product reviews contain implicit features. We counted the Chinese reviews we crawled and discovered that at least 30 percent of the sentences are implicit sentences. Meanwhile at least one implicit feature appeared in each sentence, which was a considerable proportion of our research. However, few scholars have paid close attention to implicit feature identification. Su et al. [32] used Point-wise Mutual Information (PMI), based on semantic association analysis, to identify implicit features, but no quantitative experimental results were provided. Hai et al. [14] used co-occurrence association rule mining to detect implicit features based on the opinion words; however, they seems to have neglected the facts. Poria et al. [28] present a rule-based approach that uses common-sense knowledge and sentence dependency trees to detect both implicit and explicit aspects in English reviews.

This paper proposes a semi-supervised learning approach for implicit feature identification on Chinese reviews;both the opinions and facts will be taken into account. The main idea is to generate a classifier for each product feature. Due to the fact that explicit sentences and corresponding features are drawn from the original reviews, they could be regarded as training samples to establish the classification models. Based on the different part-of-speech(POS) selection and several traditional feature selection methods, we intend to obtain different collections of training attribute to build up various vector space models (VSM). For the training model of each feature, the positive cases are the relevant explicit sentences, while the negative cases are relevant non-explicit sentences. Then several SVM classifiers are generated and applied to discriminate the non-explicit sentences.

Because the training set consists of explicit sentences, and the traditional feature selection methods cannot easily discover the training attributes which have high relevance with the product features. In fact, the training data is the explicit sentences, where each sentence contains a certain product feature and two features never co-occur in the training data. Therefore, we tend to use a clustering algorithm to gather the product feature and the relevant terms in the same topic. Topic modeling is a principled approach to solve this problem as it has the capability to gather terms of the same topic into one group. Since one cluster corresponds to one type of product feature, the terms in the same group will be highly associated. Therefore, these terms are of higher priority to be selected as the training attributes. Furthermore, topic modeling methods can be considered as clustering algorithms that cluster terms into homogeneous topics (or clusters), and the pre-existing knowledge can guide clustering algorithms to produce better and more meaningful clusters. In contrast, the basic LDA, without considering prior knowledge, cannot perform well for the following reasons:

(1) It is difficult for basic LDA to obtain the appropriate number of topics.
(2) The product features co-occur frequently with some common words such as *good*, *bad*, to name a few, which are often used to describe the product features. Therefore, when using basic LDA to cluster the explicit sentences, some product features may be clustered in the wrong topic.
(3) For each topic cluster, it is difficult to decide its corresponding product feature, which serves as a classification category of SVM.

Given the above facts, an LDA-based explicit topic model is proposed, which serves as a replacement to the traditional feature selection methods. The core idea of the explicit topic model is to pre-allocate a certain product feature to each topic. After employing certain pre-explicit knowledge concerning explicit topic features, we can obtain the corresponding category of each topic cluster based on the explicit product feature of each topic.

In our explicit topic model, two types of constraints are first extracted from the explicit sentences automatically. Then these constraints and some relevance-based prior knowledge are incorporated into the explicit topic model. Next, those explicit topic models are established to filter words and extract the training attributes of each product feature for SVM. Finally, several SVM classifiers are constructed to train the selected attributes and utilized to detect the corresponding implicit features.

The remaining parts of this paper are organized as follows. Section 2 covers some related works. Section 3 introduces the proposed approach of combining the explicit topic model with SVM to identify implicit features. In Section 4, the experimental results are reported, evaluated and discussed. Finally, Section 5 presents our conclusions.

2. Related work

In recent year, besides some innovative work in opinion mining [23], the main directions in opinion mining research include sentiment classification and feature-based opinion mining. Das and Chen [7], Dave et al. [8], Gamon et al. [12], He and Zhou [15], Ku et al. [20], Pang and Lee [24,25], Pang et al. [26], Riloff and Wiebe [31], Turney [33] who have researched on sentiment classification focus on classifying reviews of customers into three classes: positive, negative and neutral. However, only opinion reviews are classified based on the opinion holder's sentiment in their research, and the sentiment relating to product features is ignored, which make these research less valuable in practical application. In contrast, the main purpose of feature-based opinion mining is to produce a summary of reviews with different features based on sentence-level sentiment classification. Liu and Hu contribute greatly to such research field [16–18,22,21]. Meanwhile on the basis of English product reviews, they first take implicit features into account. In addition, other representative work [3,10,27] also focuses on implicit feature identification.

In this paper, we research and further utilize the definition of implicit feature which was first proposed by Liu et al. [22]. In that paper, they gave an example to show the implicit feature in a digital camera review: "*This camera is too large*". "*Size*" is an implicit feature since it does not appear in the text and is implied by the attribute word. Meanwhile, they argue that the rule-learning method can be used to generate mapping rules, and then candidate features like "heavy" are mapped to their actual features like "weight". However, they do not discuss it further and no experimental results were reported. Su et al. [32] used Point-wise Mutual Information (PMI), based on semantic association analysis, to identify implicit features, but no quantitative experimental results were provided. Hai et al. [14] used co-occurrence association rule mining to identify implicit features. However, they only dealt with opinion words and neglected factual works. Therefore, in this paper, the integration of opinions and factual words is one of the most prominent characteristics. Wang et al. [34] used hybrid association rule mining to identify implicit features, which is also a rule-based method. All these three approaches can be applied to find some uncommon but reasonable rules based on the basic rule