Information Processing and Management 51 (2015) 58–67

Contents lists available at ScienceDirect

Information Processing and Management

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

Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures

Shi Li [a], Lina Zhou [b,*], Yijun Li [c]

[a] School of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, PR China
[b] Department of Information Systems, UMBC, Baltimore, MD 21250, United States
[c] School of Management, Harbin Institute of Technology, Harbin 150001, PR China

ARTICLE INFO

Article history:
Received 28 December 2013
Received in revised form 25 July 2014
Accepted 20 August 2014
Available online 20 September 2014

Keywords:
Aspect extraction
Online reviews
PMI-IR
Frequent aspects

ABSTRACT

Online review mining has been used to help manufacturers and service providers improve their products and services, and to provide valuable support for consumer decision making. Product aspect extraction is fundamental to online review mining. This research is aimed to improve the performance of aspect extraction from online consumer reviews. To this end, we augment a frequency-based extraction method with PMI-IR, which utilizes web search in measuring the semantic similarity between aspect candidates and target entities. In addition, we extend RCut, an algorithm originally developed for text classification, to learn the threshold for selecting candidate aspects. Experiment results with Chinese online reviews show that our proposed method not only outperforms the state of the art frequency-based method for aspect extraction but also generalizes across different product domains and various data sizes.

1. Introduction

Online reviews play an important role in online advertising and marketing (Barton, 2006). Online reviews have been recognized as a type of word-of-mouth that helps managers and manufacturers in brand building, product development, and quality assurance. According to recent social commerce statistics from *Kelton Research*/*Bazaarvoice*, 83% of consumers believed that it is important to read user-generated content before making a decision about banking or other financial services. As the volume and velocity of online reviews rapidly increases, it becomes ever more challenging for customers to read through entire reviews. Aggregated ratings alone do not address the above challenge for two main reasons: (1) customers really care about aspects (i.e., parts and attributes) of a product, and (2) customers have their own preferences for aspects of a product (Liu & Seneff, 2009; Snyder & Barzilay, 2007). Therefore, product aspect extraction from the text of online reviews is instrumental for leveraging the online word-of-mouth for individual and business decision making.

Extracting product aspects from online consumer reviews remains a difficult task due to some unique characteristics of online reviews such as unstructured text, and the colloquial and casual style of Internet language (Pollach, 2005). Product reviews are commonly written by consumers, not paid professionals. The extant methods for product aspect extraction from online reviews can be classified into four main categories (Liu, 2012, chap. 5): (1) extraction based on frequent nouns and noun phrases (Hu & Liu, 2004a), (2) extraction by exploiting opinion and aspects relations (Qiu, Liu, Bu, & Chen, 2009), (3) extraction by supervised learning (Jin, Ho, & Srihari, 2009), and (4) extraction using topic modeling (Liu, Cao, Lin,

* Corresponding author. Tel.: +1 (410) 455 8628; fax: +1 (410) 455 1073.
E-mail address: zhoul@umbc.edu (L. Zhou).

http://dx.doi.org/10.1016/j.ipm.2014.08.005
0306-4573/

Huang, & Zhou, 2007). The first type of method (e.g., Apriori) is focused on those aspects that occur frequently, which depends on pruning methods to improve the relevance of the extraction results. The second type (e.g., double propagation) relies on the dependency relationship to propagate information between aspects and opinions. Although such a method is capable of dealing with infrequent aspects, noise as well as information could be propagated. The third type is most commonly used, which generally outperforms its unsupervised counterpart. Nevertheless, it requires training data, which is an inherent limitation of supervised methods. Annotating online reviews with product aspects is both labor-intensive and time-consuming, particularly in light of the high variety and volume of products and high dimensions of product aspects. The last type of method is aimed at discovering the main themes that pervade a large and otherwise unstructured collection of documents by building topic models (Blei, 2012). The topics discovered by such a method generally contain both aspects and opinions about product entities, so additional work is required to separate the two types of information.

In this research, we propose a method for product aspect extraction by augmenting frequency-based extraction with PMI-IR. PMI-IR (Turney, 2001) is used to measure the semantic similarity between aspect candidates and product entities. Compared with previous frequency-based methods, our proposed method has several advantages. First, it leverages a universal search engine rather than a static collection of online reviews to estimate the similarities between candidate aspects and an entity. A universal search engine routinely maintains the coverage and freshness of its content, which enables more updated and complete estimates of similarities with no extra effort. Second, it prunes frequent aspect candidates based on the PMI-IR score between a candidate aspect and the target entity under review instead of between the candidate and multiple discriminator phrases, which improves the efficiency of aspect extraction that uses PMI. Third, it only requires small datasets for threshold learning, which can easily scale up. Our experiment results show that the proposed methods outperform the state-of-the-art frequency-based method.

The rest of this paper is organized as follows. In Section 2, we provide background and review related work on aspect extraction. In Section 3, we introduce the experiment design in detail, followed by results and discussion in Section 4. We conclude the paper with Section 5.

2. Background and related work

The process of product review mining mainly consists of the following steps (Popescu et al., 2005): identify product aspects, identify opinions regarding product aspects, determine the polarity of opinions, and rank opinions based on their strength. Thus, product aspect extraction is fundamental to review mining. Before providing a systematic review of aspect extraction methods, we define some key concepts based on the previous literature (Hu & Liu, 2004a, 2004b; Liu, 2012, chap. 5; Liu, Wu, & Yao, 2006).

Definition (entity): Entity e is described as a target product, service, topic, issue, person, organization, or event, discussed in an online review. Entity e can be represented as a two-element vector: $e = (P, AT)$, where P denotes a hierarchy of parts which is organized based on semantic relations, and AT denotes a set of attributes of e.
Definition (aspect): Aspect $a \in (P \cup AT)$ is described as a part or an attribute of entity e.

For example, *cell phone* is the entity that a review targets, and both *screen* and *weight* are discussed in the review. Since *screen* is a part of and *weight* is an attribute of a cell phone, so both are considered as aspects of *cell phone*. Aspect has also been referred to as product feature (Hu & Liu, 2004a, 2004b) or opinion target (Jakob & Gurevych, 2010). We chose to use the term *aspect* to avoid confusion with the term *feature* used in machine learning literature (Liu, 2012, chap. 5), and to include those aspects that not have associated opinions. Aspect is commonly represented as a noun or noun phrase. There are primarily four types of methods for extracting product aspects from online reviews (Liu, 2012, chap. 5).

2.1. Extraction by frequent nouns and noun phrases

Hu and Liu (2004a, 2004b) laid the groundwork for applying Apriori algorithm of association rule mining to aspect extraction by treating frequent nouns and noun phrases as aspect candidates. The algorithm achieved a precision of 80% and a recall of 72%. Scaffidi et al. (2007) improved the above method by introducing pruning methods to filter those English words that are unlikely to be aspects, such as those words that are more likely to be a verb than a noun. In addition, they pruned some compound words based on the difference in frequency distribution between online reviews and generic text. These pruning strategies improved the precision of aspect extraction to the range of 85–90%. Nevertheless, they show two major limitations: (1) the pruning strategies were developed based on the morphological forms of English words, which are difficult to be extended to other languages like Chinese; and (2) they chose corpora of spoken and written conversations as generic text, and difference in conversational structure would have impact on the performance of text classification (Tavafi, Mehdad, Joty, Carenini, & Ng, 2013). Li, Ye, Li, and Law (2009) extended the method of Hu and Liu (2004a) to Chinese by introducing additional pruning steps aspect extraction from Chinese reviews. Xu, Huang, and Wang (2013) treated frequent sets of skip-bigrams, which are word pairs that allow skips between words, as candidate aspects. They used skip-bigrams to mitigate the negative impact of errors in Chinese part-of-speech tagging, which led to increased recall but reduced precision (below 0.5). Ferreira et al. (2008) found that Hu and Liu's approach is effective for text that mainly consists of on-topic