Knowledge-Based Systems 71 (2014) 86–100

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Product aspect extraction supervised with online domain knowledge [☆]

Tao Wang [a], Yi Cai [a,*], Ho-fung Leung [b], Raymond Y.K. Lau [c], Qing Li [d], Huaqing Min [a]

[a] School of Software Engineering, South China University of Technology, China
[b] Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong
[c] Department of Information Systems, City University of Hong Kong, Hong Kong
[d] Department of Computer Science, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Available online 20 June 2014

Keywords:
Aspect extraction
Product aspect
Topic model
Opinion mining
Review summarization

ABSTRACT

One of the most challenging problems in aspect-based opinion mining is aspect extraction, which aims to identify expressions that describe aspects of products (called aspect expressions) and categorize domain-specific synonymous expressions. Although a number of methods of aspect extraction have been proposed before, very few of them are designed to improve the interpretability of generated aspects. Existing methods either generate multiple fine-grained aspects without proper categorization or categorize semantically unrelated product aspects (e.g., by unsupervised topic modeling). In this paper, we first examine previous studies on product aspect extraction. To overcome the limitations of existing methods, two novel semi-supervised models for product aspect extraction are then proposed. More specifically, the proposed methodology first extracts seeding aspects and related terms from detailed product descriptions readily available on E-commerce websites. Next, product reviews are regrouped according to these seeding aspects so that more effective textual contexts for topic modeling are built. Finally, two novel semi-supervised topic models are developed to extract human-comprehensible product aspects. For the first proposed topic model, the Fine-grained Labeled LDA (FL-LDA), seeding aspects are applied to guide the model to discover words that are related to these seeding aspects. For the second model, the Unified Fine-grained Labeled LDA (UFL-LDA), we incorporate unlabeled documents to extend the FL-LDA model so that words related to the seeding aspects or other high-frequency words in customer reviews are extracted. Our experimental results demonstrate that the proposed methods outperform state-of-the-art methods.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The recent decade has witnessed the rapid development of E-commerce. The growth in online purchasing has resulted in a growth in online customer reviews of products. The opinions embedded in these product reviews can have a great influence on the purchasing decisions of potential customers. Typically, recommendations for a product can help increase purchases, while complaints for a product can have the opposite effect. In addition, the customer opinions can help firms develop insights into customer interests, and hence improve their products or services [1]. However, as the volume of reviews rapidly grows, it has become increasingly difficult for individuals or organizations to effectively process the wealth of information contained in the corpus of available reviews. This has given rise to the development of opinion mining techniques, which can automatically extract and summarize opinions embedded in an extensive collection of product reviews.

In order to gauge customer interests and preferences, opinions regarding specific aspects (such as, the *Lens* aspect of a camera and the *Service* aspect of a hotel) of a product are often more helpful than an overall opinion on a product. Therefore, aspect-based opinion mining, which aims at summarizing customer opinions for individual product aspects, has gained increasing attention in recent years [2–6]. In most aspect-based opinion mining systems, customer preferences are summarized based on opinions associated with specific aspects of a product. Hence, accurate aspect identification and categorization has the potential to improve the performance of opinion analysis processes. Accordingly, aspect extraction which aims at identifying expressions related to different product aspects and categorizing domain-specific synonymous expressions is an important step in opinion mining. However, the implementation of this step has been identified as a bottleneck in opinion mining systems [1,7,8].

To date, a variety of methods have been proposed for extracting product aspects from reviews. Some methods have been based on

[☆] Fully documented templates are available in the elsarticle package on CTAN.
[*] Corresponding author.

http://dx.doi.org/10.1016/j.knosys.2014.05.018
0950-7051/© 2014 Elsevier B.V. All rights reserved.

frequent-string mining [2,9–12]. However, these methods often fail to effectively categorize expressions according to their semantic content. For example, attributes of the same aspect (e.g., the attributes *"LED"* and *"LCD"* for the aspect *Screen*) are often treated as different aspects. Similarly, domain-specific synonyms describing the same aspect (e.g., *"battery"* and *"power"*) are often treated as different aspects. The opinions regarding certain aspects of a product extracted by these methods contain a certain amount of noise and are not as well-structured as they could be. As a result, it can be difficult to get a clear picture of customer opinions regarding specific characteristics of a product.

Recently, topic modeling based methods [13–17,4–6] have gained attention because they can simultaneously identify expressions that describe aspects of products (which we call *aspect expressions* in this paper) and cluster semantically related expressions. However, existing topic modeling methods have some limitations. The first limitation concerns the fact that statistically co-occurring but semantically unrelated aspects are often found in product reviews. The reason for this is that reviewers often mention multiple product aspects in a single product review. For example, the aspects *Focus* and *Lens* are typically discussed together in customer camera reviews. Existing unsupervised topic models tend to assign these frequently co-occurring terms to the same aspect (topic). This presents a serious challenge for systems that attempt to categorize semantically related aspects at the review level. An alternative approach is to perform topic modeling at the sentence level [16,17,5]. However, it is well-known that topic models perform poorly on short texts due to a lack of sufficient co-occurrence contexts [14,18]. The second limitation is that most existing topic models are unsupervised, which may cause any of a number of the problems:

- Many unrelated noise terms that have no relevance in evaluating products may be extracted by unsupervised models.
- With unsupervised models, the granularity of generated aspects (e.g., a general aspect *Screen* versus more specific aspects such as *Screen Size* and *Resolution* are representations of the same aspect at different granularities) and the associated categorization criteria may be at odds with human judgment. Thus, some generated aspects may not be readily comprehensible.
- Due to lack of supervision at the aspect-word level, the order of aspects generated by unsupervised models is usually non-deterministic, and hence may vary among different runs. Table 1 shows an example of aspects generated by unsupervised model. The *Screen* aspect was generated as the first aspect in run-1, and the second aspect in run-2. Assigning meaningful labels to such generated aspects often requires substantial manual intervention, which is tedious and time-consuming.

Since reviews always pertain to a certain product, we assume that a collection of reviews is a domain-specific corpus. Accordingly, some domain-specific expressions can be assigned to a few aspects unambiguously. The domain-specific expressions are terms which have specific meanings in a domain. An example is that the term *"apple"* may refer to a *fruit* or a *company* in a corpus of news articles. To determine the actual meaning of the term "apple", we require contextual information. For example, in reviews of the *iPhone-5*, the term *"apple"* is more likely to refer to the *company* rather than the *fruit*. We assume that some domain-specific terms in product reviews are good discriminators to indicate certain product aspects, such as *"LCD"* for *Screen* of camera. The probabilities that these discriminators can be found in the categories (or topics) that they indicate are higher than the probabilities that they can be found in other categories. Thus, domain-specific terms can be used to supervise topic models for aspect extraction.

Table 1
An example of aspects generated by unsupervised model.

Run	Aspect 1	Aspect 2	Aspect 3
1	Screen	Lens	Battery
2	Lens	Screen	Battery
...	...	...	...
n	Battery	Lens	Screen

In this paper, we extend the work presented in [4,5] on aspect extraction, and propose two semi-supervised methods by using seeding words to guide the extraction of latent product aspects from reviews. Instead of manually labeling seeding sets, as in [4,5], the proposed methodology extracts seeding set from semi-structured information about product descriptions on E-commerce websites (e.g., *Newegg.com* and *Ebay.com*). Fig. 1 shows detailed descriptions of a camera from *Newegg.com*. Expressions in each block are categorized as an aspect. The information in the product descriptions contains the primary product aspects and associated domain-specific expressions. The categorization of these product aspects, which are summarized by domain experts, is in line with human judgment. Hence, we employ the product information extracted from E-commerce websites to generate the seeding aspect set, and use it as the prior knowledge in the proposed semi-supervised topic models. The characteristics of our proposed methods and the contributions of this work are summarized as follows:

- To extract specific product aspects in line with human judgment, we employ semi-structured information about product descriptions readily available on E-commerce websites to the generate seeding sets. The seeding sets are used as prior knowledge in our proposed semi-supervised topic models for aspect extraction. Different from the manually labeling on seeding aspects in [4,5], our seeding aspects and related seeding words are automatically extracted from online domain-specific contents.
- In order to build better contexts for aspect extraction, we adopt a regrouping process which combines sentences containing the same seeding word (or the same detailed phrase in product descriptions) from different reviews into a new document. A regrouped document contains less co-occurring aspects than an original review, and has richer co-occurrence context than a single sentence. These textual contexts provide better guidance to topic models for effectively categorizing aspects. We believe that the regrouping process is an effective step to bootstrap the performance of topic models on aspect extraction. To the best of our knowledge, previously proposed methods [16,17,5,6] have not explored automatic review regrouping (regarded as a process of document context strengthening) to improve the performance of aspect extraction.
- To tackle the above mentioned limitations of unsupervised models, we have developed a semi-supervised FL-LDA to extract product aspects. By encoding the aspect-word relevance of seeding aspects, the FL-LDA can generate ordered aspects with respect to the aspects defined in the seeding set. The deterministic ordering of aspects can significantly reduce the amount of human intervention during aspect extraction; this saving may not be achieved by using previous models [16,17,5,6].
- To discover aspects that are not pre-defined in the seeding set but mentioned frequently in reviews, in the UFL-LDA, we extend the FL-LDA by incorporating sentences without seeding words. As a result, both seeding aspects and frequently occurring non-seeding aspects can be extracted.
- To demonstrate the effectiveness of the proposed models, we conducted experiments comparing the performance of our models with that of some existing aspect extraction models. The results show that both the FL-LDA model and the UFL-LDA model outperform other baseline methods.

Download English Version:

https://daneshyari.com/en/article/403596

Download Persian Version:

https://daneshyari.com/article/403596

Daneshyari.com