Title	レビューデータの意見分類に関する研究
Author(s)	Ou, Wei
Citation	
Issue Date	2018-12
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/15752
Rights	
Description	Supervisor:Huynh Nam Van, 知識科学研究科, 博士

Abstract

Nowadays, along with the rapid growth of the e-commerce economy, consumers have become more and more dependent upon review data to make decisions on shopping websites. However, reading reviews is a very time-consuming, even frustrating process when the number of reviews is overwhelmingly large. Both academia and industry have been working to devise algorithms that can automatically extract knowledge of products or services from the review data to improve users' review-reading experience. The knowledge extraction is usually treated as a sentence-level opinion classification problem, that is to detect what attributes of products or services consumers have discussed about, and how they felt of the attributes in the review sentences. This dissertation mainly focuses on the following 2 fundamental problems related to the classification task: sentence representations, the limited availability of training data. Also, based on the results of the opinion classification process, this dissertation proposes a novel machine learning based approach to identify the aspects that are generally attractive to consumers. The discovered attractive attributes allow users to quick capture the selling points of a product or service. A brief introduction to the originality of this dissertation is presented as follows.

1) Sentence representations.
Word embeddings models, as an effective way to represent text, have been widely used in various text classification tasks. Since word embeddings are only optimised to represent individual words, one has to define ways to aggregate word embeddings to represent sentences. A very effective, easy-to-compute aggregation function is averaging, though it obviously leads to loss of information. Recently, researchers have applied complex, but also computationally expensive neural network structures, such as convolutional neural network (CNN) and recursive neural network (RNN), to aggregate word embeddings. This dissertation proposes a novel weighted average approach, named `Abstract Keywords', as an alternative to the existing aggregation operators. The proposed approach assumes there exist some extremely important abstract keywords that can be derived in the training process, and assigns words different weights according to their semantic similarities to the abstract keywords. Each sentence is represented by the weighed average of the embeddings of all words in the sentence. Experiment results show that the proposed approach is computationally efficient, and outperforms the simple averaging approach.

2) Limited availability of labelled training data.
As an important aspect of review mining, sentence-level sentiment classification has received much attention from both academia and industry. Many recently developed methods, especially the ones based on deep learning models, have centred around the task. Generally speaking, training sentence-level sentiment classifiers requires training datasets of labelled sentences, that are usually every expensive to obtain. It is possible to use the less expensive labelled review documents to train sentence-level sentiment classifiers, by treating each document as a long sentence, and the label of the document as the label for the long sentence. However, this way is obviously questionable because there may exist sentences in a document whose sentiments are very different from the sentiment of the document. Therefore, the sentiments of individual sentences can be easily misrepresented by the document-level labels in the training process. To address the problem, we propose a novel approach, named `Averaged-logits', that also uses labelled documents to train sentence-level sentiment classifiers, but makes a difference by assuming different sentences in a document have different sentiments, and the `average' of the sentence-level sentiments is used to determine the document-level sentiment. In the experiment, we collected two review datasets: one contains 50,000 hotel reviews crawled from TripAdvisor, the other 50,000 electronic product reviews from Amazon. The proposed approach was evaluated on the two datasets. The results show that, the proposed approach outperforms the existing approach treating each document as a long sentence, by margins of 3\%-8\% on sentence-level sentiment classification .

3) Attractive attribute classifiers.
Researchers have proposed statistical regression models that analyse on-line review data to identify attractive attributes of a product or service. This dissertation has the same aim, but with an approach based on machine learning models instead of statistical models. The proposed approach first extracts attribute-level sentiments from the review text by natural language processing techniques, then derives features that reflect the non-linear relations between attribute performance and customer satisfaction based on the sentiments. The non-linear features are fed to the Support Vector Machine (SVM) model to train predictive attractive attribute classifiers. The proposed approach is evaluated on a hotel review dataset crawled from TripAdvisor. The experiment results indicate that the classifiers reach a precision of 79.3\% and outperform the existing statistical models by a margin of over 10\%.

Keywords: Review data; Sentiment analysis; Sentence representation; Weakly-supervised learning; Kano model analysis