氏名	KAEWPITAKKUN YONGYOS			
学位の種類	博士(情報科学)			
学位記番号	博情第 347 号			
学位授与年月日	平成 28 年 9 月 23 日			
論文題目	User-level Sentiment Analysis on Microblogging by Incorporating Target Specific Knowledge and User Network			
論文審査委員	主査	白井 清昭	北陸先端科学技術大学院大学	准教授
		東条 敏	北陸先端科学技術大学院大学	教授
		Nguyen Minh Le	北陸先端科学技術大学院大学	准教授
		池田 心	北陸先端科学技術大学院大学	准教授
		藤井 敦	東京工業大学	准教授

論文の内容の要旨

Microblogging services have been becoming increasingly popular over the last decade. Many people express their opinions and feeling about anything in the famous microblogging service, Twitter. These people's opinions can be grabbed easily and publicly through the interface provided by Twitter. Both individuals and organization are increasingly using this data for decision making. Customers want to know the opinion of other users before making purchase decision. Companies want to know the feedback from users about their products and also their competitors. Therefore, opinion mining and sentiment analysis become one of the major research topics in the field of natural language processing. Early work on the sentiment analysis proposed methods of classifying the sentiment on the traditional social network, i.e. forum, webboard and review. However, it is more difficult to analyze sentiments on tweets. Tweets are very short and contain a lot of informal expressions, i.e. slang, emoticon, typographical errors and a lot of words that are not compiled in a dictionary. The solution and method of traditional sentiment analysis system cannot be applied directly because of these unique characteristic of the microblogging. Moreover, existing sentiment analysis approaches mainly focus on measuring the sentiment of individual tweets or predict the massive opinions for a specific target. However, identification of the opinions of individual users is another important task that is often required in practical opinion mining systems.

In this thesis, we research and develop several methods of classifying the sentiments on microblogging, aiming to address the problems described above. We do not only focus on classifying the sentiment of each tweet by considering the textual information, which is usually short and hard to interpret. We aim to seek other characteristics in microblogging to extract the extra knowledge for boosting the performance of the sentiment analysis. Three main sentiment analysis tasks are considered, namely tweet-level sentiment analysis, target-dependent sentiment analysis, and user-level sentiment analysis. In the following, we describe each analysis one-by-one.

First, in the tweet-level sentiment analysis, we introduce a hybrid approach that uses a lexicon for sentiment words to alleviate the data sparseness problem inherent in machine learning approaches and improve sentiment classification in tweets. The data sparseness problem can be reduced by the following two methods. We first estimate the potential polarity of objective and out-of-vocabulary (OOV) words and use these words as additional information of the existing sentiment lexicon. The polarity scores of OOV words are estimated based on the assumption that the polarities of words are coincident with the polarity of their associated sentences, using a collection of labeled sentences with their polarity. Then, we introduce a novel feature weighting method by interpolating sentiment lexicon score into uni-gram score in the feature vectors of SVM.

Second, in the target-dependent sentiment analysis, we propose a method for incorporating on-target sentiment information and user sentiment information into a machine learning classifier for the target-dependent sentiment analysis of the tweets. Three extra resources, the add-on lexicon, the extended target list, and the competitors list, are automatically constructed from the unlabeled tweets. The target specific training data is created based on heuristic rules and the lexicon-based sentiment analysis method. Two new features for training the sentiment classifier are introduced. One is the on-target sentiment feature, giving greater weight to the sentiments of the words near the target; the other is the user sentiment feature that captures the tendency of the sentiment expressed by the same user.

Third, in the user-level sentiment analysis, we propose a novel graph-based method that incorporates the information of both textual information, as well as the explicit and implicit relationships between the users, into a heterogeneous factor graph for the sentiment analysis of the tweets at the user level. Our framework takes into consideration not only the explicit connections such as follow, mention and retweet but also the implicit connections between users. An implicit connection refers to the relations of users who share similar topics of interest. The implicit relations among the users are extracted from their historical tweet corpus. Since the presence of the explicit relations in some social network is limited, the implicit relations allow us to utilize the data in social network more effectively. We also propose a new enhanced pooling method, "Hashtag-PMI", to more precisely infer the latent topics by the conventional LDA (Latent Dirichlet Allocation) from the tweet corpus.

Both public and real-life tweet corpora are used in our experiments. The results of experiments show that our method achieves 64-70%, 59-62% and around 65% accuracy on the tweet-level, target dependent and user-level sentiment analysis tasks respectively. The proposed method is effective and significantly improves the performance compared to the several baselines and existing methods.

Keywords: Sentiment Analysis, Machine Learning, Knowledge Acquisition, Topic Modeling, Microblogging

論文審査の結果の要旨

本論文は、マイクロブログのユーザが特定の対象に対して表明した意見の極性(肯定的か否定的か)を判定する新しい手法を提案している。ユーザが投稿したテキスト(ツイート)から得られる情報に加え、ユーザ間の明示的な関係、暗黙的な関係の情報を利用し、極性判定の性能を向上させている点に特長がある。

まず、ツイートの極性を判定する手法を提案した。既存の感情語辞書が、ツイートに頻出する口語的表現を十分にカバーしていないという問題に対し、ツイートの集合から感情語の辞書を自動構築する手法を提案した。また、自動構築された感情語辞書を機械学習モデルに組み込むことでツイートの極性判定の正解率を向上させた。

次に、ツイート内で表明された特定の対象(製品、人物、企業など)に対する意見の極性を判定する手法を提案した。この手法では、まず、対象に特化した知識として、前述の方法で構築された感情語辞書、対象表現リスト、競合実体リストを作成した。次に、これらの知識といくつかのヒューリスティクスにより、対象の極性判定に特化した訓練データを自動構築した。最後に、この訓練データから、特定の対象に対する意見の極性を判定するモデルを機械学習した。この際、対象と感情語の距離を考慮した素性と、同一ユーザによる別のツイートの極性を考慮した素性を新たに提案した。

最後に、特定の対象に対するユーザの極性を判定する手法を提案した。提案手法は以下の 3 つの情報を利用する。1 つ目は、ツイート上に表明されている意見の極性である。これは前述の特定の対象に対する極性を判定するモデルを用いて推定した。2 つ目はユーザ間の明示的な関係である。本研究では retweet 関係を利用した。3 つ目はユーザ間の暗黙的な関係である。似たトピックに興味を示すユーザは同じ極性を示す可能性が高いという仮定の下、これをユーザ間の暗黙的関係と定義し、過去のツイートを集積したデータから暗黙的関係を自動的に推定する手法を提案した。さらに、これら 3 つの情報を総合的に判定してユーザの極性を推定するグラフベースの手法を提案した。実験の結果、対象に特化した極性判定モデルがグラフベースの手法において有効に働くこと、ユーザの暗黙的関係を新たに導入することによりユーザの極性判定の正解率が向上することを確認した。

以上、本論文は、マイクロブログを対象とした感情分析に関する新しい手法を提案し、優れた成果を示したものであり、学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。