Title	他者のコメントの引用を考慮したオピニオンマイニング
Author(s)	岡山，有希
Citation	
Issue Date	2013-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/11340
Rights	
Description	Supervisor:白井清昭，情報科学研究科，修士

Opinion Mining Considering Citations of Others' Comments

Yuki Okayama(110015)

Japan Advanced Institute of Science and Technology School of Information Science

February 06, 2013

キーワード: Opinion Mining, Polarity, Citation of comment, Sentiment Analysis.

In recent years, a large amount of texts can be easily obtained from World Wide Web. Text mining is an important technique to get knowledge from a huge collection of documents. On the other hand, people often express their opinion in blogs or microblogging. Among texts on the Web, articles in blogs are valuable resources to know emotions and sentiments of people. Nowadays, opinion mining, a kind of text mining, is noteworthy as an important technology to know people's opinion, and blog are an useful knowledge sources for opinion mining. However,we should pay attention that bloggers (authors of blogs) often cite other blogs or news articles. In some cases, the blogger expresses a positive opinion while the cited text expresses a negative opinion, and vice versa. It may cause misclassification of polarity of the blog article. The goal of this paper is to develop an opinion mining system as follows: for a given topic, it retrieves the related blog articles from the Web, detects and removes the cited text from the article, classifies if the article expresses positive, negative or neutral opinion to the topic and show the number of articles for each category(positive,negative,neutral). It also shows users the sentences indicating positive and negative opinions. Such a system enable us to overlook opinions of people in the world about the topic. Past studies of opinion mining have not been devoted to handle the cited text. The main contribution of this thesis is to propose a method considering the cited texts for classification of the blog article as positive or negative.

The system consists of three tasks and seven modules. Three tasks are 'Comment Retrieval' task, 'Citation Detection' task, and 'Polarity Judgement' task. Seven modules are 'Blog Search','Sentence Segmentation','Citation Detection/Removal', 'Sentence Selection', 'Sentence Polarity Judgement', 'Blog Polarity Judgement' and 'Counting' module. The goal of Comment Retrieval task is to retrieve the blog articles related to the given topic form the Web. Blog Search module searches blogs that contain query keywords. Sentence Segmentation module first identifies the main contents (not advertisement, table of

contents of the blog site and so on) by heuristic, then splits it into sentences by HTML tags and punctuation. The goal the second Citation Detection task is to remove the cited texts which are quoted from other websites. Citation Detection/Removal module finds citation blocks and citation sentences. Citation block is defined as a node in Document Object Model (DOM) of the web page, which covers cited texts. First, 308 keywords indicating citation (called citation keyword) are prepared. Nodes near the citation keyword are detected as citation nodes by several heuristics. While, each sentence in the blog page is judged if it is cited comment or not in order to find citations that citation keywords do not appear near. The linked web pages are used for judgement of citation sentences. The similarity between the sentences of the blog page and the sentences of the linked web page are calculated. The system determines the sentence as the cited sentence if it can find a sentence whose similarity is high enough in the linked web page. The similarity between two sentences are defined by Levenshtein distance (Edit Distance). The goal of the third Polarity Judgement task is to identify the polarity of the blog article. In this task, first Sentence Selection module selects the sentences that are supposed to express the opinions of the blogger. They are called 'opinion sentence candidates' in this study. Sentences near all query keywords are chosen as opinion sentence candidates. Then, chosen candidate sentences are classified as positive, negative or neutral by Sentence Polarity Judgement module. This module uses evaluative verbs and nouns in the Evaluative Expression Dictionary for polarity classification. Next, Blog Polarity Judgement module classifies the polarity of the blog article based on the results of polarity of opinion sentence candidates. Finally, Counting module accumulates number of the blog articles classified as positive or negative. This module counts numbers in two ways: 'total count' and 'unique count'. The former is the number of positive or negative blog articles when the polarity is judged by all the sentences, while the latter is one when the polarity is judged by only sentences except for citation.

The proposed methods are evaluated by the experiments. Six queries are used for evaluation, which are sets of keywords related to the topic (ex."baby hatch, criticism","olympic, Tokyo, place"). For each query, top 50 blogs are retrieved by the search engine "Yahoo! Search Blog". Then the performances of Citation Detection, Sentence Polarity Judgement and Blog Polarity Judgement are evaluated on the retrieved blog articles. In order to evaluate the citation detection, sentences in the evaluation data are manually judged if they are cited comments or not. Precision and recall and F-measure of citation detection were 94%, 71% and 81%, respectively. Precision was relatively high, but 71% recall was still low. One of the reasons is that the system fails to detect citation blocks since there is no citation keyword. Another reason is that the boundary between the author's text and citation text is unclear. Next, the polarity judgement of the sentences are evaluated. In this experiment, only sentences where the system judges as positive or negative are evaluated. In other words, precision is measured, but recall is not. Two evaluation criteria are used. 'Accuracy' is the ratio of the correct sentences, where the gold and predicted polarities are agreed and the sentences express the opinions for the topic given by the query. While 'Accuracy (polarity)' is the ratio of the sentences, where polarities are agreed but the sentences express the opinions for other topics (not the query topic). In the experiments, Accuracy was 18%, While Accuracy (polarity) was 42% . There were many errors that neutral sentences are misclassified as positive or negative since there exists evaluative expressions (especially nouns) in the sentence. Finally, polarity judgement of blogs is evaluated. Three systems are compared. System1 determines the polarity using all sentences in the blog article. System2 determines the polarity after cited texts are removed. In this system, cited texts are automatically detected and removed by the proposed method. System3 also determines the polarity without cited texts, but cited texts are removed by hand. Accuracy of polarity judgement of system1, system2 and system3 was 55%, 58% and 60%, respectively. Therefore, accuracy of the polarity judgement was improved by removing cited texts. Then, F-measure for each polarity category is evaluated. F-measure of positive category of system 1, 2 and 3 was 19%, 18% and 18%. F-measures of negative category of system 1, 2 and 3 were 43%. F-measure of neutral category of system 1, 2 and 3 was 70%, 73% and 75%. We found that the neutral blogs were classified the most correctly.

In this paper, the method of opinion mining that takes into account cited texts is proposed. Experimental results show that handing cited text is important for the opinion mining system when blog articles are used as knowledge source. However, the performance of polarity judgement is still poor. There is much room for improvement. For instance, the use of dependency relations in the sentence would be helpful for the polarity judgement. The practical evaluation to see how useful the proposed system is for overlooking people's opinions is another important future work.