Automatic Classification of Link Polarity in Blog Entries

Aya Ishino, Hidetsugu Nanba, and Toshiyuki Takezawa

Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194, Japan
{ishino,nanba,takezawa}@ls.info.hiroshima-cu.ac.jp

Abstract. In this paper, we propose a method for classification of an author's sentiment for a linked blog (we call this sentiment link polarity), as a first step for finding authoritative blogs in the blogosphere. Generally, blogs that are linked positively from many other blogs are considered more reliable. In citing a blog entry, there are passages where the author describes his/her sentiments about a linked blog (which we call citing areas). We extract citing areas in a Japanese blog entry automatically, and then classify a link polarity using the information in the citing areas. To investigate the effectiveness of our method, we conducted experiments. For classification of link polarity, we obtained a high precision and recall than baseline methods. For the extraction of the citing areas, we obtained the same Precision and Recall as manual extraction. From our experimental results, we confirmed the effectiveness of our methods.

Keywords: Blog, Link polarity, Sentiment analysis.

1 Introduction

Recently, with the explosive spread of blogs, users can express their private ideas or opinions on the internet easily and actively. The importance of this information is recognized widely; however, the information in blogs contains a mixture of wheat and chaff. Therefore, identifying reliable information efficiently has become an important issue. Many researchers have been trying to collect individual opinions from blogs and analyze them. We have been studying the automatic identification of authoritative blogs in the blogosphere.

At present, methods based on the number of links are used to find authoritative information, such as the PageRank algorithm used by Google [1]. However, the algorithm does not reflect author sentiment about the site being linked. Therefore, blogs on the receiving end of abuse are sometimes highly ranked on by search engines.

To solve this problem, we focus on the author's sentiment about a linked blog. We call this sentiment link polarity. In a citing blog entry, there are passages where the author describes their sentiments about a cited blog, as shown in Figure 1 (which we call citing areas). In the case of Figure 1, both cited blogs A and B have three citing blogs. Based on the number of links, these cited blogs have the same authority. However, we identify cited blog A as more authoritative using link polarity.

M.V.M. Salem et al. (Eds.): AIRS 2011, LNCS 7097, pp. 509–518, 2011.

In this paper, we automatically extract the citing areas in a blog entry, and classify the link polarity using the information within the citing areas. This information is useful for identifying authoritative blogs in the blogosphere efficiently, because blogs that are linked positively from many other blogs are considered more reliable.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 explains our methods for classifying link polarity. To investigate the effectiveness of our methods, we conducted some experiments, and Section 4 reports the experimental results. We present conclusions in Section 5.

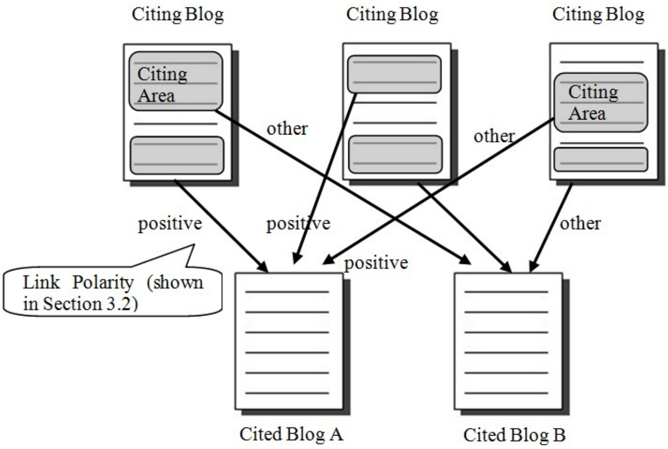

Fig. 1. Citation relationships between blogs

2 Related Work

In this section, we describe some related studies on sentiment analysis, lexicon for sentiment analysis, and classification using citing area.

2.1 Sentiment Analysis

In this paper, we focus on the author's sentiment for a linked blog, classifying sentiment as a link polarity. We can regard the classification of link polarity as a kind of sentiment analysis. Nanno *et al.* [2] presented a system called "blogWatcher", which collects Japanese blogs, performs searches on them, and classifies the sentiment for a search query as positive, negative or other, using a sentiment analysis technique. If a user uses the name of a commercial product, such as "iPhone", as a search term, the system extracts and classifies the sentiment for "iPhone" as a polarity for the commercial product.

There have been several reports of research that automatically classify online movie reviews [3, 4]. Turney applied an unsupervised learning technique based on polar words, such as "excellent" and "poor" [3]. Manual information is computed using statistics gathered by search engines. Pang *et al.* classified reviews using the three machine learning methods (Naive Bayes, maximum entropy classification, and SVM) [4]. They used uni-gram and bi-gram as features for machine learning.

Nanno *et al.* classify the polarity of commercial products. Turney and Pang *et al.* classify movie reviews. In our work, we aim to classify the polarity of author sentiment of a cited blog.

2.2 Lexicon for Sentiment Analysis

For sentiment analysis, researchers use a lexicon that contains polar phrases and their polarity values. Kaji and Kitsuregawa built a lexicon for sentiment analysis from Japanese HTML documents [5]. We call the lexicon the sent lexicon. The sent lexicon contains approximately 10,000 Japanese polar phrases with their polarity values. The polarity value is a numerical value indicating the strength of polarity, which is referred to as a polarity value. To disambiguate orthography, all polar phrases were transferred into their original form with Juman[1] software for morphological analysis. For example, the sent lexicon contains 綺麗だ ("beautiful"), but does not contain either きれいだ or キレイだ, which translate to the same word. When blog authors mention the link, they sometimes use お勧め ("recommend") and参考 ("reference"). As the sent lexicon contains adjectives and adjective phrases, these words are not included. In this paper, we build a lexicon for classification of link polarity. We call this lexicon the link lexicon.

2.3 Classification of Link Polarity

There have been several reports on research to classify links in blog entries automatically [6, 7]. Kale *et al.* devised a method that classifies links in blog entries as either positive or negative [6]. They used a lexicon of positive and negative oriented words and matched the token words from *x* characters before and after the links to determine the polarity. The link polarity was calculated as follows.

$$Polarity=(N_p\text{-}N_n)/\ (N_p+N_n) \qquad (1)$$

Here, N_p is the number of positively oriented words and N_n is the number of negatively oriented words. We manually created rules for the automatic extraction of any surrounding sentences that mention the link (citing areas). Alternatively, Martineau proposed a machine-learning approach for link classification from several viewpoints using words that appear in the context of citations of URLs as features [7].

Several researchers focused on links in a social net [8, 9]. Guha and Kumar studied data sets from a general consumer review site, Epinions [8]. Leskovec *et al.* studied data sets from Epinions, Slashdot, and Wikipedia [9]. They predicted positive and negative links in data sets using a method based on machine learning, and compared these with theories of balance and status from social psychology. They use the

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

number of links as a feature for machine learning. In this paper, we focus on links in blog entries, and classify the link polarities using the sentiment in the text surrounding the link.

In a research paper, there are passages where the author describes the essence of a citing paper and the differences between the current paper and the cited paper (we call them citing areas). These passages can be considered as a kind of summary of the citing paper from the current author's viewpoint. Nanba *et al.* [10] automatically classified the types of citation relationships that indicate the reason for citation, using information in the citing areas. They classified the reason for citation into three categories.

To construct collections of hyperlinks for a tourist spot, Ishino *et al.* [11] classified the link in travel blog entries [12] into four categories with citing areas. They manually created rules using cue phrases for the automatic extraction of citing areas. For the classification of links, they obtained a high precision with the information in the citing areas.

In this paper, we automatically extract citing areas in the same way as Nanba *et al.* and Ishino *et al.*, and classify link polarity with the information in the citing areas.

3 Classification of Link Polarity

In this section, we define link polarity in Subsection 3.1, and then explain our approach for classification of link polarity in blog entries in Subsection 3.2.

3.1 Link Polarity

Authors of blogs often link to other blogs with a sentiment for a linked blog. We call the sentiment link polarity. In general, a link polarity can be classified into three categories: positive, negative and other. However, there were only five negative links in 840 links that we collected. Because the link polarities were overwhelmingly positive, we classify link polarities in this paper as either positive or other.

3.2 Classification of Link Polarity

The procedure for the classification of link polarity is as follows.

1. Input a blog entry.
2. Extract a hyperlink and any surrounding sentences that mention the link (a citing area).
3. Classify the link by taking account of the information in the citing area.

In the following, we will explain Steps 2 and 3.

Extraction of Citing Areas. We manually created rules for the automatic extraction of citing areas. These rules use cue phrases. When authors of a blog mention links, they use particular words, such as "blog" (ブログ), "entry" (記事) , or author's name/title of the linked blog. Therefore, we manually selected the cues, and used them for citing area extraction. For extraction of citing areas, we take account of the HTML anchor types and classify the anchor into the following three categories.

- Anchor type 1: Anchor is a URL of linked blog.
- Anchor type 2: Anchor contains blog author's name of linked blog.
- Anchor type 3: Other than anchor types 1 and 2.

We manually created rules 1, 2, and 3, for Anchor types 1, 2, and 3, respectively.

- Rule 1 for anchor type 1
    1. Extract a sentence that includes the link and extract *X* sentences that appear before or after a web hyperlink, and add them to the candidate. Here, we used the value of $X = 2$, which was determined via a pilot study.
    2. Extract the author's name or title of the linked site from one sentence that appears before or after a web hyperlink. When blog authors introduce web sites, quotation marks or brackets are often used immediately before and after the title of the site. We extract character strings within quotation marks or brackets as keywords. We also extract a word with "Mr." (君) and "Ms./Mrs." (さん) as an author name of the linked blog.
    3. Extract all sentences including the title of the linked blog in the blog entry as citing areas.
    4. Extract all sentences including the author's name of the linked blog in the blog entry as citing areas. If cues appear in a sentence that appears before or after the sentence including cues, we extract them as citing areas.
- Rule 2 for anchor type 2
    1. Extract a word with "Mr." (君) and "Ms./Mrs." (さん) as the author's name of the linked blog.
    2. Extract all sentences including the author's name of the linked blog in the blog entry as citing areas. If cues appear in a sentence that appears before or after the sentence including the title of the linked blog, we extract them as citing areas.
- Rule 3 for anchor type 3
    1. Extract a sentence that includes the link and extract *X* sentences that appear before or after a web hyperlink, and add them to the candidate. Here, we used the value of $X = 2$, which was determined via a pilot study.
    2. Extract sentences include the same character strings as the anchor and character strings just before particular cues, such as "blog of" (のブログ), or "entry of" (の記事), as the author's name of the linked blog.
    3. Extract all sentences including the author's name of the linked site in the blog entry as citing areas. If cues appear in a sentence that appears before or after the sentence including cues, we extract them as citing areas.

Building the Link Lexicon for Classification of Link Polarity. We build a lexicon containing positively oriented words for classification of link polarity. We call the lexicon the link lexicon. If the citing area contains positively oriented words, we classify the link polarity as positive. If not, we classify the link polarity as other. The procedure for building the link lexicon is as follows.

First, we collect sentences containing the word “this blog” (このブログ) as candidates for the polar sentence from blog entries and a “Web Japanese N-gram” database[2] provided by Google, and manually classify the polarity of the sentences. Examples of polar sentences are shown in Table 1. Second, we manually extract positively oriented words from sentences classified as positive. The double-underlined parts are the positively oriented words. We extract positively oriented words, such as “recommended” (オススメ) and “filled with information” (情報満載). Finally, we record the positively oriented words in the link lexicon. The link lexicon contains 135 positively oriented words.

Table 1. Examples of polar sentences

Polarity	Polar sentence
positive	**[original]** このブログはマジでオススメである。 このブログは情報満載なのでリンクさせて頂きます。 **[translation]** This blog is strongly recommended. I link to this information-rich blog site.
other	**[original]** このブログは、オレの気の向くままに運営しております。 このブログは参加しにくい。 **[translation]** I administer this blog as my fancy dictates. It is difficult to participate in this blog.

4 Experiments

To investigate the effectiveness of our methods, we conducted several experiments.

4.1 Data Sets and Experimental Setting

We randomly selected 840 links in blog entries, manually classified the link polarity, and used them for our examination. Table 2 shows the number of positive links and other links. We used precision, recall and F-measure as evaluation measures.

Table 2. The number of hyperlinks for each type

Link polarity	Positive	Other	Total
The number of links	378	462	840

[2] http://www.gsk.or.jp/catalog/GSK2007-C/catalog.html

In this paper, we propose two methods for building the link lexicon for classification of link polarity and extraction of citing areas. Thus, we considered and evaluated two different approaches. First, to evaluate the link lexicon, we use blog data A with manual extraction of citing areas. Second, to evaluate the method for extraction of citing areas, we use blog data B with automatic extraction of citing areas by our method.

4.2 Evaluation of the Link Lexicon

To evaluate the link lexicon, we use blog data A with manual extraction of citing areas manually. There are two main methods.

[Methods based on Kale *et al.*'s words]

- Kale_LinkLex (our method): Use the link lexicon.
- Kale_SentLex[θ]: Use the sent lexicon. By using polarity value and threshold θ(>0), we decided whether a word is a polar word or not. If the polarity value of the words is greater than θ, we regard the word as a positive word. Similarly, if the polarity value of the words is less than $-\theta$, we regard the word as a negative word. We calculate link polarity using equation (1).

[Methods based on Martineau's method]

We used the TinySVM (http://chasen.org/~taku/software/TinySVM/) software as the machine-learning package. We performed a four-fold cross-validation test.

- Martineau_LinkLex (our method): Use the link lexicon as features.
- Martineau_SentLex (our method): Use the sent lexicon as features.
- Martineau_Base: Use the words as features.

Results and Discussion

The evaluation results are shown in Table 3. Kale_SentLex[θ] achieved a high precision with large θ. Kale_SentLex[θ] achieved a high recall with low θ. We found Kale_SentLex[11] achieved the highest precision and Kale_SentLex[3] achieved the highest recall, as shown in Table 3. Kale_LinkLex obtained a high precision and recall in comparison with the baseline method based on polar words. Finally, Martineau_LinkLex obtained a high precision and recall in comparison with the baseline method based on Martineau's method. The methods using the link lexicon obtained the highest precision and recall. Therefore, we have confirmed the effectiveness of the link lexicon that we built for classification of link polarity. In addition, Kale_LinkLex obtained the best performance.

Table 3. Evaluation results for link lexicon

		Precision (%)	Recall (%)	F-measure (%)
Methods based on Kale *et al.*'s method	**Kale_LinkLex (our method)**	**85.2**	**90.3**	**87.8**
	Kale_SentLex[11]	72.0	9.5	40.8
	Kale_SentLex[3]	55.9	67.7	61.8
Methods based on Martineau's method	**Martineau_LinkLex (our method)**	**81.7**	**74.8**	**78.3**
	Martineau_SentLex (our method)	**78.8**	**74.1**	**76.5**
	Martineau_Base	78.0	71.6	74.8

The performances of methods based on Kale's method were better than the performances of methods based on machine learning. When blog authors introduce a linked blog, they use particular words, such as "recommended" (オススメ) and "nice" (ステキ). There are a few variations of polar words in blogs. Therefore, we obtained higher performances with methods based on Kale *et al.*'s method than Martineau's method. We show the number of correct and incorrect classification in Table 4. There are many differences in the incorrect number between Kale_LinkLex and Martineau_LinkLex. If we combine the methods based on Kale *et al.*'s and Martineau's methods, and we can further improve the performance of link polarity classification.

Table 4. Evaluation results of Kale_LinkLex and Martineau_LinkLex

		Martineau_LinkLex		Total
		correct	incorrect	
Kale_LinkLex	correct	640	105	745
	incorrect	44	51	95
Total		684	156	840

With Kale_LinkLex, there were two typical errors in the classification of link polarity: (1) the lack of polar words in the link lexicon and (2) the limitation of the method depending on polar words. We describe these errors as follows.

(1) Lack of cues

For classification of link polarity, we manually collected polar words and built the link lexicon, as described in Subsection 3.3. To improve the coverage of polar words, a statistical approach, such as applying n-gram statistics to a larger blog corpus, will be required.

(2) Limitation of the method depending on polar words
Our method mistakenly classified the following example as positive. We proposed the method based on taking account of polar words in citing areas for classification of link polarity. In the example, “nice” (ステキ), which was collected as a polar word, appears in the citing areas, so our method classified it as positive. The blog author said that the present was nice. However, the blog author did not say the linked blog is nice. Our method cannot analyze what is nice. To improve the performance of classification of link polarity, we need to consider language structure.

[original]
先日いつも仲良くしていただいている
<a href=“http://blogs.yahoo.co.jp/tenmomomini/” target=_blank>テンファミリー＋オチビ</a>の「もりりんさん」から、　ステキなプレゼントが届きました〜〜
[translation]
The other day, I received a nice present from Ms. Momorin who was the blog author of <a href=“http://blogs.yahoo.co.jp/tenmomomini/” target=_blank>Tenfamily and Ochibi</a>

Fig. 2. An example of a failure

4.3 Evaluation of Citing Areas

To evaluate our methods for extraction of citing areas, we use blog data B with automatic extraction of citing areas by our method.

- Kale_LinkLex_Auto (Our method): Use the link lexicon as features.

Results and Discussion
The evaluation results are shown in Table 5. Kale_LinkLex_Auto used blog data B with automatic extraction of citing areas by our method. Kale_LinkLex used blog data A with manual extraction of citing areas. The result of Kale_LinkLex is from Table 3. Kale_LinkLex_Auto shows the same precision and recall as Kale_LinkLex. Therefore, we have confirmed the effectiveness of our method for the extraction of citing areas. In this paper, we proposed the method for extraction of citing areas using cues. In our future work, we will consider linguistic information from the linked site and sentences surrounding the link and improve the performance of extraction of citing areas.

Table 5. Evaluation results for citing areas

	Precision (%)	Recall (%)	F-measure (%)
Kale_LinkLex_Auto (our method)	**86.5**	**87.6**	**87.1**
Kale_LinkLex (our method)	**85.2**	**90.3**	**87.8**

5 Conclusion

In this paper, we proposed methods for classification of link polarity in blogs. First, we collected polar words and built the link lexicon. Next, we extracted citing areas from blog entries and classified the link polarity in the blog as positive or other using the link lexicon. In the evaluation of the link lexicon, Kale_LinkLex obtained precision and recall scores of 85.2% and 90.3%. Martineau_LinkLex obtained precision and recall scores of 81.7% and 74.8%. Kale_LinkLex obtained the best performance. In the evaluation of our method for the extraction of citing areas, Kale_LinkLex_Auto shows the same precision and recall as Kale_LinkLex with manual extraction of citing areas. The experimental results confirmed the effectiveness of our methods.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems Archive 30(1-7), 107–117 (1998)
2. Nanno, T., Fujiki, T., Suzuki, Y., Okumura, M.: Automatically Collecting, Monitoring, and Mining Japanese Weblogs. In: 13th International World Wide Web Conference, Poster Session (2004)
3. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424 (2002)
4. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Conference on Empirical Methods in Natural Language Processing, pp. 76–86 (2002)
5. Kaji, N., Kitsuregawa, M.: Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In: Conference on Empirical Methods in Natural Language Processing (EMNLP-CoNLL 2007), pp. 1075–1083 (2007)
6. Kale, A., Karandikar, A., Kolari, P., Java, A., Finin, T., Joshi, A.: Modeling Trust and Influence in the Blogosphere Using Link Polarity. In: International Conference on Weblogs and Social Media (2007)
7. Martineau, J., Hurst, M.: Blog Link Classification. In: International Conference on Weblogs and Social Media (2008)
8. Guha, R.V., Kumar, R., Raghavan, P., Tomkins, A.: Propagation of Trust and Distrust. In: 13th WWW (2004)
9. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting Positive and Negative Links in Online Social Networks. In: 10th WWW (2010)
10. Nanba, H., Okumura, M.: Towards Multi-paper Summarization Using Reference Information. In: 16th International Joint Conferences on Artificial Intelligence, pp. 926–931 (1999)
11. Ishino, A., Nanba, H., Takezawa, T.: Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries. In: ENTER 2011 (2011)
12. Nanba, H., Taguma, H., Ozaki, T., Kobayashi, D., Ishino, A., Takezawa, T.: Automatic Compilation of Travel Information from Automatically Identified Travel Blogs. In: The Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, Short Paper, pp. 205–208 (2009)