Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events

Tomohiro FUKUHARA
The University of Tokyo
5-1-5 Kashiwanoha, Kashiwa,
Chiba 277-8568 JAPAN
+81-4-7136-4275
fukuhara@race.u-tokyo.ac.jp

Hiroshi NAKAGAWA
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku,
Tokyo 113-0033, JAPAN
+81-3-5841-2729
nakagawa@dl.itc.u-tokyo.ac.jp

Toyoaki NISHIDA
Kyoto University
Yoshida-Honmachi, Sakyo-ku,
Kyoto 606-8501 JAPAN
+81-75-753-5371
nishida@i.kyoto-u.ac.jp

ABSTRACT

Temporal sentiment analysis that analyzes temporal trends of sentiments and topics from a text archive that has timestamps is proposed. The method accepts texts with timestamp such as Weblog and news articles, and produces two kinds of graphs, i.e., (1) *topic graph* that shows temporal change of topics associated with a sentiment, and (2) *sentiment graph* that shows temporal change of sentiments associated with a topic. Sample results obtained by applying the method to news articles are described.

Keywords

Temporal sentiment analysis, topic detection

1. INTRODUCTION

Today, understanding concerns of people is important in various domains. Among such concerns, sentiment of people is important because people's sentiment has great influence on our society. The aim of this research is to analyze trends of people's sentiment along with timeline. Figure 1 illustrates an overview of the proposed method called *temporal sentiment analysis*. The method takes texts with timestamps such as Weblog and news articles as input, and creates two types of graphs, i.e., 1) *topic graph* which shows topical trends associated with a specific sentiment (e.g., anxiety), and 2) *sentiment graph* which shows trends of sentiments associated with a topic (e.g., earthquake). Because users can specify sentiment and a topic, s/he can easily find trends of topics and sentiments.

2. PREVIOUS WORK

In the natural language processing communities, many works on sentiment analysis are conducted. Turney proposed a method for assessing semantic orientation (SO) of a text[1]. Pang et al. proposed a method for assessing positive or negative of a sentence using several machine learning techniques[2]. These researches, however, focused on binary evaluation, i.e., positive or negative. In this research, we focus on multi-polarity sentiment analysis.

ICWSM'2007 Boulder, Colorado, USA

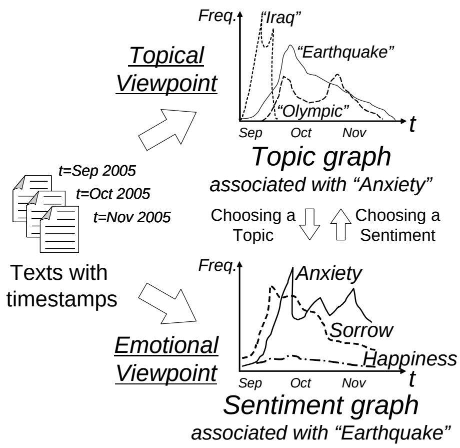

Figure 1: Overview of the temporal sentiment analysis.

Mishne and de Rijke proposed a system called *MoodViews*[1] that can analyze temporal change of sentiment[3]. MoodViews analyzes multiple sentiments by using 132 sentiments used in LiveJournal[2]. Although our concept for the sentiment graph is similar to MoodViews, we also focus on temporal change of topics associated with a sentiment in the topic graph.

With respect to topic detection and tracking (TDT) researches, several methods are proposed for detecting and tracking events from text archives[4]. However, few researches consider the utilization of sentiment in the TDT domain. We focus on *sentiment biased* topic detection.

With respect to information visualization, Havre et al. proposed a system called *ThemeRiver*[5] that visualizes thematic flows along with timeline. Although our approach is similar to ThemeRiver, we focus on visualization of both of topical and sentiment flows along with timeline.

3. TEMPORAL SENTIMENT ANALYSIS

We describe an overview of the temporal sentiment analysis.

3.1 Overview

The method produces topic graph and sentiment graph by using *sentiment phrases* which are patterns of sentiment expression such as "happy" or "delighted at". We extracted 383 sentiment phrases

[1] http://moodviews.com/

[2] http://www.livejournal.com/

from Japanese news articles manually, and classified them into eight categories: *anxiety*, *sorrow*, *anger*, *happiness*, *suffering*, *fatigue*, *complaint*, and *shock*.

3.2 Procedure for Making a Topic Graph

Following is the procedure for making a topic graph.

Given: one of sentiment category S which is specified by a user
period of time: $D=(d_1, d_2, \ldots, d_l)$

Step 1: For each day d_i in D, retrieve articles containing sentiment phrases of sentiment s.

Step 2: Extract keywords from retrieved articles by using a keyword extraction system called *GENSEN-Web*[3] that can extract compound nouns as a keyword[6].

Step 3: For each extracted keywords $w_j(j=1,2,\ldots,N)$, calculate an average correlation c between w_j and sentiment phrases contained in S. We use the Dice coefficient for calculating correlation[4].

Step 4: Extract top n keywords according to the score defined by the products of (1) number of days in which keywords appears, (2) inverse frequency of number of days, and (3) scores provided by GENSEN-Web.

Step 4'(optional): Put keywords into clusters based on correlation coefficient over timeline and the Dice coefficient in an article.

Step 5: Generate a temporal graph for each n keywords (or clusters). For viewability of the graph, we apply moving average.

3.3 Procedure for Making a Sentiment Graph

Following is the procedure for making a sentiment graph.

Given: a keyword w which is specified by a user
period of time: $D=(d_1, d_2, \ldots, d_l)$

Step 1: Retrieve articles containing keyword w for each day $d_i(i=1,2,\ldots,l)$.

Step 2: For each articles, calculate the sum of frequency of sentiment phrases for all sentiment categories.

Step 3: Generate a temporal graph of frequency of sentiment phrases for each sentiment category. Then, moving average is applied to the graph.

4. SAMPLE RESULTS

Sample results obtained from the proposed method are shown in Figure 2 and Figure 3. We applied our method to a news corpus from Asahi Shimbun newspaper[5], which is written in Japanese, published in 2004. Figure 2 shows an example of a topic graph associated with "happy" in 2004. x axis is the date, and y axis is the dice coefficient between sentiment phrases and keywords. We can see topics associated with "happy" such as the homecoming of abduction victims by North Korea.

Figure 3 shows an example of a sentiment graph associated with "earthquake" in the fourth quarter in 2004. x axis is the date, and y axis is frequency of sentiment phrases. This graph is a stacked chart. We can see bursts of "shock" and "anxiety" when the earthquake occurred (around 22 Oct).

[3] http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_eng.html

[4] We chose the Dice coefficient because it showed a good performance among other coefficient functions.

[5] Total number of articles in the corpus is 15,455.

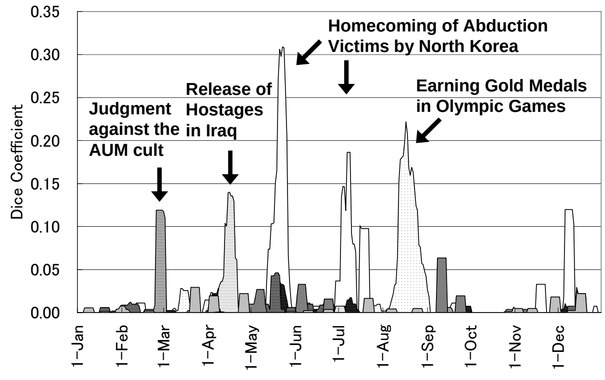

Figure 2: Topic graph for sentiment "happy" in 2004 (using clustering option).

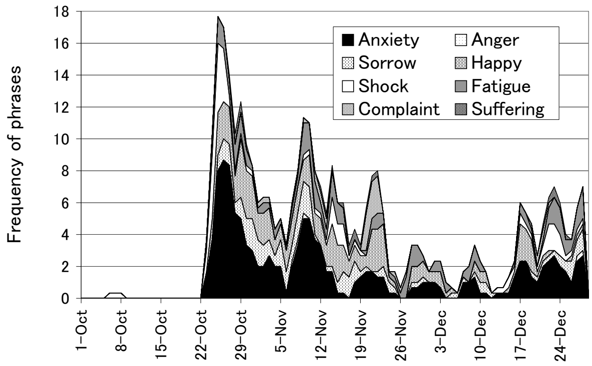

Figure 3: Sentiment graph for the topic "earthquake" in the fourth quarter in 2004. (stacked chart)

5. CONCLUSION

We proposed a method for analyzing temporal trends of sentiments and topics from texts with timestamps. Our future work contains (1) qualitative and quantitative evaluation, (2) automatic construction of a database of sentiment phrases, and (3) to apply the proposed method to Weblog articles.

REFERENCES

[1] Turney, P.D., Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, proceedings of the 40th ACL, pp.417-424, 2002.

[2] Pang, B. Lee, L., and Vaithyanathan, S., Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of EMNLP2002, pp.79-86, 2002.

[3] Mishne, G. and de Rijke, M., MoodViews: Tools for Blog Mood Analysis, AAAI 2006 Spring Symposium on Computational Approaches to analyzing Weblogs (AAAI-CAAW2006), 2006.

[4] Wayne, C.L., Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, Language Resources and Evaluation Conference (LREC), pp.1487-1494, 2000.

[5] Havre, S., Hetzler, E., Whitney, P., and Nowell, L., ThemeRiver: Visualizing Thematic Changes in Large Document Collections, IEEE Transactions on Visualization and Computer Graphics, Vol.8, No.1, pp. 9-20, 2002.

[6] Nakagawa, H. and Mori, T., Automatic Term Recognition based on Statistics of Compound Nouns and their Components. Terminology, Vol.9 No.2, pp.201-209, 2003.