D-045

Real-Time Support Rate Estimation Based on Social Sensors

Huang Jun† junhuang@akane.waseda.jp

Mizuho Iwaihara† iwaihara@waseda.jp

ABSTRACT

This paper attempts to propose estimation of real-time support rate based on social sensors. Nowadays, micro blog like *Twitter* has gained wide popularity and therefore received much attention academically. Personal opinions are posted online by Twitter users frequently, especially when big events with worldwide concern occur. In the case that follows, each Twitter user is treated as a sensor and therefore collectively, the global Twitter users are referred to as Social Sensors. Drawing on one soccer game, we collect a large amount of tweets and carry out analysis so as to extract sentiment information of the audience and show the real-time support rate of the two teams.

1 . INTRODUCTION

As a popular micro-blogging service provider, *Twitter* has grasped tremendous attention from public. People can update their status anywhere in the form of *tweet* which is a message within 140 characters with the aid of computer or mobile phone. Therefore, millions of people are benefited from this social network service that enables them to make new friends and keep connection with their friends, classmates or colleagues.

Topic of a tweet can be variable. Mostly, tweets help people express their opinions of a certain topic, like a new product, a game or presidential campaign. By extracting the sentiment information of such tweets, support rate for each aspect (percentage of positive tweets to certain aspect) of the topic can be calculated. These sentiment information and support rate are rather important since feedbacks are aggregated without manual intervention.

Moreover, given *Twitter*'s real-time characteristic, any change of the support rate affected by the sudden presence of a big event, even the most trivial one, can be caught immediately, like a goal during a game or a speech given by a president candidate during election debate.

1.1 Related work

Enormous efforts have been made in the area of sentiment analysis [2]. Traditionally, methods mainly focus on static corpus, like reviews, yet not without flaws. Due to the real-time characteristic of *Twitter*, the number of tweets centering on one specific topic could grow at an astonishing pace. So far, peek value of number of tweets per second was 3283 during the game between Japan and Denmark in World Cup 2010. Lacking the ability of reflecting continuous real-time change of the sentiment information (support rate) for a great deal of tweets is the major disadvantage of conventional methods.

Contributions of this paper are summarized as follows:

1. The support rate calculation is based on real-time data collected from Twitter. It inherits the distinctive feature of real-timeness;
2. A real-time event detection mechanism based on social sensors [1] is deployed in our approach, thus the change of support rate caused by this event can be caught promptly;
3. With the real-time support rate and real-time event detection, prediction for future support rate will be easier.

The rest of the paper is organized as follows: data preparation is demonstrated in Section 2. Section 3 illustrates the approach of extracting sentiment information and calculating the real-time support rate using machine learning method such as Support Vector Machine (SVM). The final part is devoted to the discussion of prospective improvements of present work.

2. DATA PREPARATION

We developed a web Twitter mining tool by using Twitter API[1] . After setting query words, a request will be sent to the server of Twitter, then a response with 15 entries in an atom file will be supplied as input to the analysis program. With the limitation of connections to Twitter, about 2250 latest tweets talking about a given topic can be collected every hour.

2.1 Tweets

Each tweet contains 12 attributes among which only 4 will be examined: published time, title, language and geo information if it is not empty. Although the attribute "content" also covers information contained in attribute "title", it would not be used since it involves html information.

Though only tweets in English are considered in our system, support rate distribution categorized under languages remains one of the most promising areas of future improvement regarding this estimation.

2.2 Topics

Topics that are chosen should be sentiment-oriented. A soccer match would be a good candidate because there exist multiple possibilities of happenings that may influence the event and furthermore affect the support rate. For example a goal, a beautiful shot or an unexpected winner could largely influence the attitude of the audience.

Recently UEFA2010 has finished in May and World Cup 2010 was just kicked off in South Africa. We gathered tweets focusing on these matches. During each match, the names of the two teams are set as keywords for query and over 5000 tweets are collected for one match.

3. APPROACH

The aim of our approach is to calculate real-time support rate and show the reflection of events which may to a large extent change the attitude of audience.

3.1 Preprocess

Usually language used in a tweet is informal. Words like "Goooooooooal" or "Yeahhhhhhh" are very common, yet such

†Graduate School of Information, Production and Systems, Waseda University

1. http://apiwiki.twitter.com/Twitter-API-Documentation

repeated characters should be removed. Also, letters "h", "w" and "l" occurred repeatedly in the tail of a word need to be removed. After preprocessing there will be no more than two consecutive occurrence of a letter. Examples of the preprocessing are "Gooal" and "Yeahh".

3.2 Training Data

Tweets collected during UEFA Champion League Final are used for training. 200 tweets are picked up manually and each of them is attached with certain sentiment values. Here we define the sentiment value of certain tweet is "a personal positive or negative feeling to the query word".

3.3 Event Detection

At the early stage of our estimation emphasis will be mainly put on detecting an event like a goal, including finding the information about which team scores, when and who.

Tweets that describe a goal can be extracted and classified into three features:

"Goal": Presence of "Goal" is the most important feature of this event;

A score report: "x-x", "x vs. x" or "x to x";

Player name (optional): Name of the player who just scores.

Although by detecting the presence of these features in a tweet a goal can be confirmed, checking whether the next several tweets also describe a goal is necessary to avoid false positives.

3.4 Support Vector Machine

Support Vector Machine is a popular classification method. By preparing positive and negative tweets as training data, it automatically produces a classifier and classifies tweets into 2 categories (positive and negative). LibSVM[2] is used in our experiment.

Data used in SVM should be vectors instead of natural languages. Feature values are extracted from tweets and then vectors are generated as input to SVM. Table 1 shows twelve features used in our estimation:

Feature	Value Type
Tweet Length	integer
Number of occurrence of first query word	integer
Position of first occurrence of first query word	integer
Position of last occurrence of first query word	integer
Number of occurrence of second query word	integer
Position of first occurrence of second query word	integer
Position of last occurrence of second query word	integer
Number of occurrence of players of first team	integer
Nmber of occurrence of players of second team	integer
Event symbol, such as "Goal"," score"	integer
Current scores of first team, if mentioned	integer
Current scores of second team, if mentioned	integer

Table1. Features used in SVM

For judging whether a feature is present, we first split tweets into tokens. Then a particular phrase comparison is required. Considering casual language used in Twitter and frequent misspelling, edit distance between two given phrases should be calculated. If it is less than two, presence of the current feature can be confirmed. Since the chance of misspelling occurring in the first and last letter in a word is tiny, a higher weight is set to these letters while calculating edit distance.

2. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

3.5 Experiment

In this section we use tweets collected during the match of England against Germany in World Cup 2010 to conduct our experiment. Reflection of the support rate of events like a goal is clearly recognized. Fig 1 shows real-time support rate of England during the period from 26' to 52' of the game.

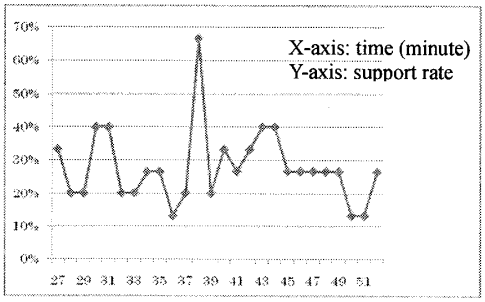

Fig1. Real-time support rate of England

Results generated by the classifier trained with SVM are reasonable. Support rate of England was relatively lower than Germany when the latter took one goal lead at the beginning. It experienced fluctuation for a few minutes which however, was followed by a sudden drop when Podolski made Germany's second goal at 32'. Yet, a significant rise of England's support rate occurred when they dramatically scored two goals in just one minute. But the high support rate of almost 70% did not last long as the second goal was disallowed and the following tweets could not be classified into a positive category since most of them were discussing the mistake made by the referees.

4. CONCLUSION AND FUTURE WORK

Machine learning can help to extract the polarities of tweets with a satisfactory accuracy. Nevertheless since only events like a goal are detected, others such as a beautiful shot are ignored. In future work, we consider detecting more varieties of events using social sensors. Also, the following aspects are expected to help improve the accuracy of support rate.

Language and geo distribution: sentiment distribution categorized by language and geo information is required for future work which is of central importance in building multi-dimensional distribution of support rate.

Domain independence: cross-domain classifier will make current estimation much more practical. Potential of current work in the business scene can be explored

Neutral tweets: semantic analysis to neutral tweets may help to improve the accuracy of current estimation.

Community support rate: with a deep mining of the relation between tweets and retweets, there is a high possibility of extracting the polarity of a community to a given aspect.

REFERENCES

[1]Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors in WWW2010, April 26-30, 2010, Raleigh, North Carolina.

[2] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang and Zheng Chen. Cross-Domain Sentiment Classification via Spectral Feature Alignment in WWW2010, April 26-30, 2010, Raleigh, North Carolina.