Twitter Trends Detection by Identifying Grammatical Relations

Mikhail Dykov, Pavel Vorobkalov

Volgograd State Technical University, Volgograd, Russia
dmawork@mail.ru, pavor84@gmail.com

Abstract

The problem considered in this paper relates to identification of trends in a given area based on analysis of Twitter messages. The approaches currently used for Twitter trends detection are based on n-grams. We propose another approach of trend detection based on identifying trend as grammatical relation and perform the identification of trending relations on the basis of their frequency change dynamics. This paper describes our method, which evaluates grammatical relations in a flow of messages on a particular subject taking into consideration both their frequency and semantic similarity among the pairs of relations. We conducted experiments to compare the outcomes provided by our method with the trends detected by conventional Twitter algorithms. The results confirmed the effectiveness of our method. The trends identified from the application of our method are easier for human interpretation.

Introduction

Twitter is a very popular micro-blogging service. Now it has more than 200 million registered users (Song, Li and Bao 2012). Twitter’s users generate about 50M tweets per day.

One of the Twitter’s features is “Trending Topics” - words and phrases, highlighted on the main page of Twitter, that are currently popular in users’ tweets. Trending topics are identified for the previous hour, day and week. Trends usually consist of breaking news, emerging events, and general topics that attract the attention of a large fraction of Twitter users. Trend detection is important for online marketing professionals and opinion tracking companies, as trends point to topics that capture the public’s attention (Mathioudakis and Koudas 2010). Determining trending topics can be considered a type of First Story Detection (FSD), a subset of the larger problem known as Topic Detection and Tracking (Allan, Papka and Lavrenko 1998). In this paper we propose a method, which determines trend in Twitter by identifying Grammatical Relations. Under the term "Grammatical Relation" we understand the edge in the Stanford dependency tree, the syntactic relationship between two words present in a clause: the head and its dependent. For example, the sentence: "She gave me a raise" contains grammatical relation "(gave, raise)". The type of this relation is "direct object". The direct object of a verb phrase is the noun phrase which is the object (accusative) of the verb (Marneffe and Manning, 2008). Our approach allows extraction of trends in a form easy for human interpretation.

Related Work

The popularity of Twitter presents many challenges for applications of Natural Language Processing (NLP) and machine learning (Benhardus 2010). Twitter has been used to study the dynamics of social networks, particularly the temporal behavior of social networks (Perera, Anand, Subbalakshmi and Chandramouli 2010). First story detection (Petrovic, Osborne and Lavrenko 2010), trend and event detection techniques (Glance, Hurst and Tomokiyo 2004) and data mining for Twitter trending topics summarization have also been applied to Twitter (Sharifi, Hutton and Kalita 2010).

Twitter has its own algorithm for generating a list of main topics (Cheong and Lee 2009). Michael Mathioudakis and Nick Koudas developed TwitterMonitor, which performs real-time trend detection, based on identification of ‘bursty’ keywords (Mathioudakis and Koudas 2010). James Benhardus presented an approach of streaming trends detection based on the TF-IDF metric calculation (Benhardus 2010).

All these approaches are n-gram based. They have some limitations. Unigram approaches can't identify trends, containing several words not merged in a hash tag, if some of the words are high-frequency and some are low-frequency. For example, the recent two words trend: "New music". The words "New" and "music" would never appear in trends separately, because they have a stable frequency within the timeline.

The approaches based on bigrams, trigrams and bigger n-grams have one common problem, i.e. they do not detect trends at early stages of their formation, because multiword trends appear not so frequently as single word trends. It is also hard to detect trends for which the community has not developed a stable set and order of key words. For example, the following phrases, taken from tweets: "Obama won presidential elections" and "Obama won elections" have the same meaning, but, using n-grams approach, they would be divided into two separate trends.

The available approaches also fail to account for the semantic similarity among words in different messages, which leads to some information loss during the trend analysis. For example, the following two tweets have the same meaning: "The democrats won in the American elections" and "Democrats won the popular vote", but this can't be detected using above mentioned approaches.

The approach which we suggest allows identifying grammar relations from messages, which makes it possible to single out and identify trends formulated by means of a few key words with different word orders. Our approach also accounts for the semantic similarity among relations, which makes it possible to identify emerging trends against the background of information noise and more accurately estimate the popularity of trends described with different words in different messages.

Methodology

Twitter provides API which allows extracting messages relating to given key words for the last 7 days and metadata of their authors. To accelerate the analysis, we decided to extract 100 random messages for every hour within one week. The manual analysis of the most frequent words, which were used by users, in the extracted messages has shown that the users who have the smallest number of followers often use spam words. We also found that there is almost no difference in the frequency of words used by all users and the users having the largest number of followers. That is why we decided to exclude from the further analysis the messages of those users who were among the lowest 10% in the number of followers.

We performed POS tagging, using Stanford Parser (Marneffe, MacCartney and Manning 2006), to clip non-informative words, as well as stop words for each message. We considered words relating to the following parts of speech: nouns, adjectives, and verbs, as they are the most significant for trend detection, because most of the words in Twitter trends and in trending hash tags belong to these parts of speech. For our approach we extract only grammatical relations with words belonging to one of the selected parts of speech.

We performed the identification of trending relations on the basis of their frequency change dynamics, which allows screening out relations containing words which are not present in frequency dictionaries and not related to the area being considered.

Input and Preprocessing

To identify trends, it is necessary to have a sample from Twitter N messages over a 7-day period:

$$S = \{M_1, M_2 ... M_N\} \quad (1)$$

$$M_i = (D, T, U) \quad (2)$$

where M_i is a message present in the Twitter search results on the targeted key words, D is the date of message creation, T is the text of the message, and U is the author of the message.

$$T = (W, H, Ul, L, R) \quad (3)$$

where W is a set of message words, H refers to hash tags, Ul is a set of user references, L is a set of hyperlinks, and R is a set of user identifiers who retweeted the message.

$$U = (Id, F) \quad (4)$$

where Id is a user identifier and F is the logarithm of the number of followers.

$$\begin{gathered} W = \{c_1, c_2, ..., c_k\} \\ c_i = (w, p, sen) \\ H = \{\#c_1, \#c_2, ..., \#c_{k2}\} \\ Ul = \{@c_1, @c_2, ..., @c_{k3}\} \end{gathered} \quad (5)$$

where c_i is a concept, w is a word, p is a part of speech, and sen is a sense. p is a noun, a verb, or an adjective. The labeled parts of speech will be used later in identifying semantic similarity among words using WordNet.

Relations Extraction and Scoring

Let us define the set of key grammatical relations of the subject area as:

$$RL = \{(rl_1, r_1), (rl_2, r_2) ... (rl_t, r_t)\} \quad (6)$$

where rl_i is the grammatical relation, r_i is the relation rating in $\{W_1, W_2 .. W_n\}$.

$$rl = \{RT, O_1, O_2\} \quad (7)$$

where RT is the relation type, O_1, O_2 - are objects(words) included in the relations. The following relation types are key in trends analysis: dependent, direct object, clausal complement with external subject, adjective modifier. Thus,

$$RT \in \{\text{dep, conj, dobj, xcomp, amod, nn}\} \quad (8)$$

where dep - dependent relation, conj - conjunct relation, dobj - direct object relation, xcomp - clausal complement with external subject relation, amod - adjective modifier relation, nn - noun compound modifier.

We use Stanford grammatical relations notation and Stanford parser to singled them out (Marneffe and Manning, 2008). Relation is considered if $(O_1 \in C)$,and $(O_2 \in C)$, where C is a set of all possible concepts for the tweet message.

To determine the rl_1 relation rating, we propose the following method:

$$r_i = a_i + \sum_{j=1, j \neq i}^{t} k_{ji} \cdot a_j \tag{9}$$

where a_i is a number of entries of rl_i relations to $\{W_1, W_2 .. W_n\}$, k_{ji} is the semantic closeness ration between rl_i and rl_j relations.

$$k_{ij} = \begin{cases} 0 \quad if \quad (RT_i \neq RT_j) \cup \left(\overline{synonym(O_i, O_j)} \cap \overline{hyperonym(O_i, O_j)}\right) \\ \dfrac{1}{d(O1_i, O1_j) + d(O2_i, O2_j)} \end{cases} \tag{10}$$

where d is a distance between concepts in WordNet (Miller, Beckwith, Fellbaum, Gross and Miller 1993).

To minimize the time needed for finding semantic closeness for each word pair, we indexed all possible synonyms and hyperonyms for all $\{O_1, O_2 .. O_N\}$.

$$In = \{In_1, In_2 ... In_M\}$$

$$In_i = \{\{O_j, d(O_i, O_j)\}, \{O_{j+1}, d(O_i, O_{j+1})\} .... \{O_{j+k}, d(O_i, O_{j+k})\}\} \tag{11}$$

$$synonym(O_i, O_j) \cap hyperonym(O_i, O_j) \cap (O_i = O_j)$$

where In is an index of synonyms and hyperonyms.

Thus, for identifying words $\{O_i, O_{i+1} .. O_{i+k}\}$ that have a certain semantic closeness with word O_j and distances d to these words, it is necessary to find $In_m, O_j \in In_m$ and single out all links related to In_m from the index.

Identification of Trending Relations

During the calculation of the relations rating over a certain period, some general high frequency relations which are not related to the theme of the messages may be found among the top query results of a given query. To eliminate this kind of problem, we decided to take into consideration the change dynamics of the relations rating instead of the rating itself. The application of this method makes it possible to exclude general high-frequency relations, as their frequency has a permanent distribution in time. Thus, the final relations dynamic rating is calculated in accordance with the following formula:

$$r_T = \begin{cases} \dfrac{r_{D+t}}{r_D + k} \quad if \quad r_{D+t} => r_D \\ \dfrac{-r_D}{r_{D+t} + k} \quad if \quad r_{D+t} < r_D \end{cases} \tag{12}$$

where r_T is the final relations rating, r_{D+t} is the relations rating for the period from D to $D+t$, r_D is the relations rating for the period from $D-t$ to D, and k is the change weight reducing ration for low rating relations which depends on the messages total count.

Thus, the relations with the greatest final rating will be trends pertaining to the topic of interest.

Experiments and Results

We decided to apply our method on the set of Twitter messages that were generated during the USA president elections which took place on Tuesday, November 6, 2012. Using Twitter API and the keyword "elections" we obtained 10.000 messages from the 5th till the 9th of November. We also obtained information about the count of the followers of the messages' authors. The creation dates of Tweets were evenly distributed within the period. First, we sorted messages by the authors' followers count and then took the first 90% of the messages. This was done to remove spam messages. Next we did some preprocessing work: removed retweets, '#' and '@' signs, URLs. Then we extracted all selected parts of speech and grammatical relations from the rest of the messages using Stanford Parser. For our experiments we decided to divide messages into groups by their creation time by 3 ways: by hours, by 6 hours and by days increments. After, we used our method (9) for each group of messages and for each selected period to calculate weights of all extracted relations independently. After that, we recalculated weights based on relation usage/frequency change dynamics in the messages groups within the time periods (12). Finally, we compared the results with trends, generated using TF-IDF based unigram and bigram approaches described by Benhardus (Benhardus 2010). For unigrams and bigrams we used words, extracted during POS tagging. We also recognized auxiliary verbs as stop words for bigrams.

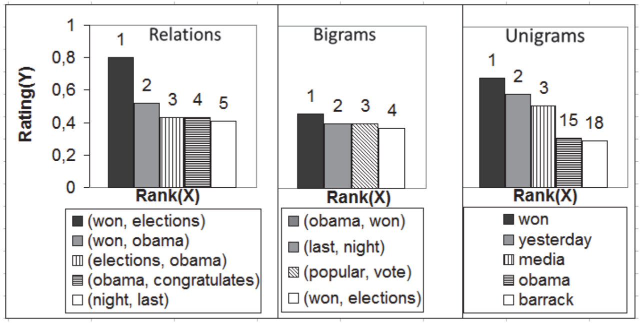

Figure 1. Top trends selected by our method, unigram and bigram approaches 6 hours after elections.

The numbers above chats describe the rank of relations, unigrams and bigrams in the trends rank list.

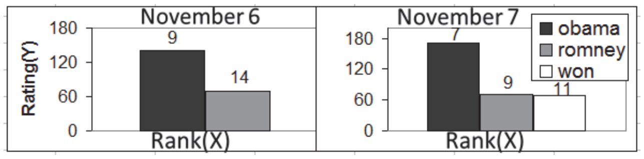

Figure 2. Top unigrams used in Tweets on November 6 and 7.

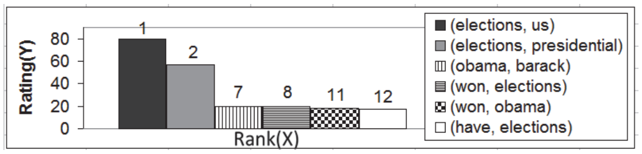

Figure 3. Top Relations used in Tweets on November 7.

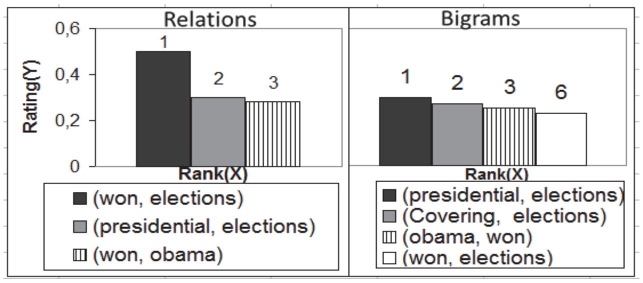

Figure 4. Top Relations and Bigrams used in Tweets 1 hour after the end of elections.

The Y axis on Figures 1 and 4 shows the rating, calculated with formula (12) for relations and for bigrams and with formula (9) for unigrams. On Figures 2 and 3 rating is equal to the frequencies of words or relations.

Discussion and Conclusion

From Figure 1 we can see that in the unigram approach there is a very large difference between ratings and positions in the trends rank list among "obama" and "won" trends. Therefore, it is hard to determine who is actually won the elections. Figure 2 demonstrates the reason: the frequency of the word "obama" stays the same in each selected time period during the elections. From Figure 1 we can see that our approach shows more “human understandable” results than the unigram approach - it is definitely clear that there were elections and Obama won. The bigram approach also shows good result on the selected test set. But, there is a big difference between the rating of bigrams "(obama, won)" and "(won, elections)" and between the rating of relations "(won, elections)" and "(won, obama)". That occured because many "Obama won elections" tweets contain the following phrases: "Obama really won the presidential elections", "won the American elections". This means, that the bigram approach can not well detect trends with separated words, especially, as we can see on Figure 4, at early stages of their formation. The high rating of "(won, elections)") trend also caused by the fact that some tweets contained phrases like "won the popular vote" instead of "won the elections". Calculation of semantic closeness with formula (10) between relations "(won, elections)" and "(won, vote)" added an additional rating to each relation. Figure 3 shows, that if we use relations ratings just among a certain time period, we get relations that contain common words: common for current event ("presidential", "elections") and common everywhere ("have"). As we can see, this problem was solved by the calculation of the dynamic of the relation usage instead of static statistic over the period. But now our approach has some limitations:

- It is not so fast for whole tweets stream processing as aforesaid bigram or unigram approaches, because POS tagging, relations extraction and the semantic closeness calculation takes time. But our method can still be used in real-time for keyword based trends extraction.
- It doesn't detect very well trends at the early stages of their formation, when not many tweets were processed. It happens because the selected POS tagger and the relations parser don't show high accuracy on Twitter messages. But in the most cases there are enough tweets that can be recognized by the relations parser for trends detection, and still, our method shows better results on this stage than the bigram approach.

References

Song, S.; Li, Q.; and Bao, H. 2012. Detecting Dynamic Association among Twitter Topics. *WWW 2012*: 605-606.

Mathioudakis, M.; Koudas, N. 2010. TwitterMonitor: trend detection over the twitter stream. *ACM:* 1155-1158.

Allan, J.; Papka, R.; and Lavrenko, V. 1998. On-line New Event Detection and Tracking. *ACM SIGR*: 37-45.

Marneffe, M. and D. Manning. 2008. The Stanford typed dependencies representation. *COLING 2008*.

Benhardus, J. 2010. Streaming Trend Detection in Twitter. *UCCS REU*.

Perera, R.; Anand, S.; Subbalakshmi, P.; and Chandramouli, R. 2010. Twitter Analytics: Architecture, Tools and Analysis. *MILITARY COMMUNICATIONS CONFERENCE 2010*: 2186-2191.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming First Story Detection with appilcation to Twitter. *NAACL:*181-189 .

Glance, N.; Hurst, M.; and Tomokiyo, T. 2004. Blogpulse: Automated Trend Discovery for Weblogs. *WWW 2004 Workshop*.

Sharifi, B.; Hutton, M.; and Kalita, J. 2010. Experiments in Microblog Summarization. *NAACL-HLT 2010*.

Cheong, M.; Li, V. 2009. Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base. *CIKM 2009 Co-Located Workshops*:1-8.

Marneffe, M. MacCartney, B. and D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. *LREC*.

Miller, G. Beckwith, R. Fellbaum, C. Gross, D. and K. Miller. 1993. Introduction to WordNet: An On-line Lexical Database.