Twitter Search Methods using Retweet Information

Jaeyoung Chang
Department of Computer Engineering
Hansung University
Seoul, Korea
jychang@hansung.ac.kr

Han-joon Kim
School of Electrical and Computer Engineering
University of Seoul
Seoul, Korea
khj@uos.ac.kr

Abstract— **Recently, as social network services such as *Twitter* and *FaceBook* are becoming more popular, a large number of researches have been carried out with various approaches. However, since social network services have been launched recently, its related search methods are still at an early stage of practical service. Thus, most of current web search sites provide a simple search service for social network service posting articles in the order of their upload time. In this paper, we present a novel way of searching informative posting data in *Twitter*. The proposed method uses both the frequency of *retweets* and the number of users' followers as major factors of ranking function in order to evaluate the quality of postings.**

Keywords-Twitter; social network service; ranking; search

I. Introduction

Currently, Web 2.0 provides a variety of advanced services such as information sharing, information generation and user-friendly service. One of the popular Web 2.0 services is the social network service such as *Twitter* and *FaceBook*. Especially, the *Twitter* service begun in 2006, and the number of its members increases exponentially and is now over 200 millions.

The posting articles of Twitter are different from the general Web contents in two aspects. First, the posting articles (shortly, *tweet*) are limited to be 140 bytes in length; thus, the Twitter users should compose their thoughts and information concisely and clearly. Secondly, the tweets can be disseminated extremely fast due to the *retweet* and *mention*; the *mention* allows Twitter users to deliver their opinions for each tweet in a question/answer form. The *retweet* is to automatically transmit the tweets to users' *followers*, and thus it allows users to share significant tweets with the followers. As a result, twitter users can read their followers' tweets in nearly real time.

However, in order to obtain new and valuable information in Twitter, users have only to depend upon their followers, even though more than 200 million tweets in every single day are spread out in Twitter social network [1]. Thus, twitter users come to expect some effective search methods in tweets as in Web document search. Of course, many of Twitter web pages provide a sort of search service, but they show only the tweets containing given search keywords in the order of posting time. By contrast, the current Web search engines can find out Web documents relevant to given keywords by using geometric (or semantic) similarity functions and analyzing hyperlink structures, and they show the search results in the order of similarity values.

The approach to searching tweets is very different from the one to searching Web documents. Since tweets are very short, it is not easy to evaluate the relevance of the articles only with search keywords. In addition, many of tweets contain users' sentiments or opinions, and so such subjective articles (or sentences) should be excluded for search service. With considering these characteristics, one better solution to searching tweets is to use their meta-information rather than their contents; the meta-information includes the number of a user's followers, the frequency of *retweets*, the frequency of *mentions*, link information, and so on.

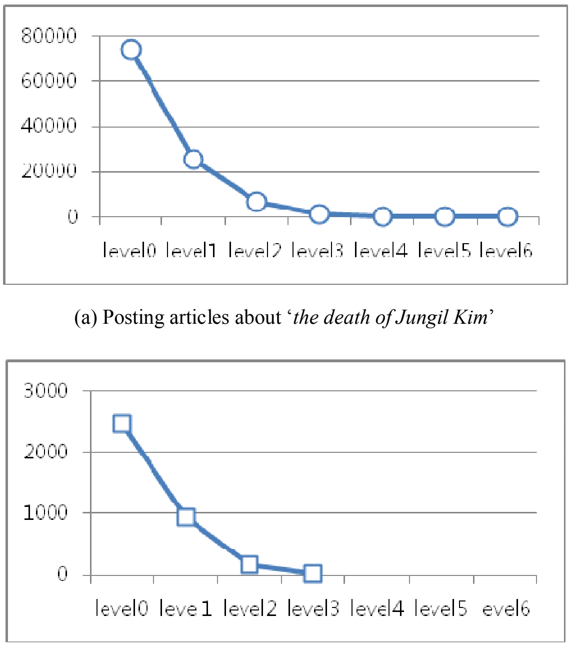
This paper proposes a novel way of searching tweets utilizing the *retweet* meta-information. Recent research work in [2-4] tackled the search problem for tweets, and however it is still at an early stage of practical service. Particular tweets to be *retweeted* are informative ones that users have judged to be worthy of dissemination. Thus the *retweet* frequency is very useful for evaluating the importance of initial tweets. The proposed method considers not only the number of followers but also the retweet frequency.

The rest of this paper is organized as follows. In Section II, we review several related research work on analyzing social network data. In Section III, we describe interesting and useful patterns for the number of followers and *retweet* frequency. In Section IV, we present three search ranking functions based on the observed patterns. In the last section, we summarize our work and introduce future work.

II. Related Work

The conventional TF-IDF (Term Frequency–Inverse Document Frequency) based search methods are not suitable for search methods of tweets [5-6]. It is because the TF-IDF weighing scheme is effective for the case that the term frequency is a good indicator for the importance of documents. Since a tweet normally does not have more than two significant words, it is very hard for TF and IDF to be metrics suitable for tweets.

Until now, several Twitter search methods have been proposed, which are usually based on the meta-information. A. Sarma, At. Sarma, S. Gollapudi, and R. Panigrahy [3] proposed a ranking method that exploits users' feedback information, and its ranking function was implemented by analyzing the users' feedback score data. However, this method is dependent all upon users' feedback, and thus it is



(a) Posting articles about '*the death of Jungil Kim*'

(b) Posting articles about '*the service shut down of a Korean broadcaster*'

Figure 1 The frequency of *retweet* operations

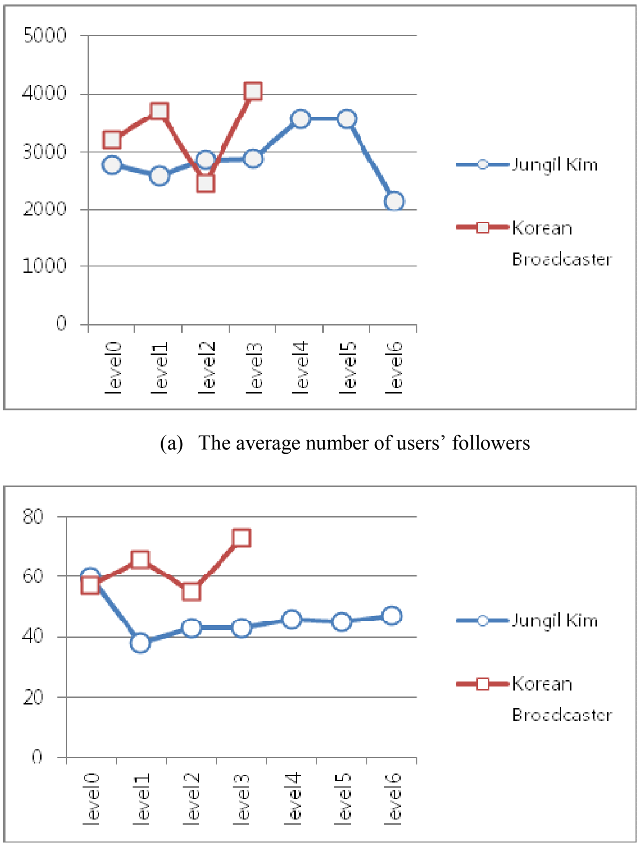
impractical to search tweets. H. W. Lauw, A. Ntoulas, and K. Kenthapadi [4] proposed another search method that actively uses meta-information such as the *retweet* frequency, the number of followers, and the posting time interval. This method is more or less similar to our proposed method. However because it estimates the *retweet* frequency only with the number of followers, the estimated retweet frequency can be incorrect in many cases. Our proposed method tries to use the precise *retweet* frequency. Also, R. Nagmoti and M. D. Cock [2] proposed a ranking method that uses the number of tweets, the number of followers, and the length of tweets.

Actually, most of studies on Twitter have focused upon searching influencing users instead of searching interested tweets [7-11]; for example, in [10], the Google's PageRank method was applied to the Twitter service in order to evaluate users' effect for each other. However, we believe that future systems will require their sophisticated search methods for social network data.

As mentioned before, some previous studies consider the *retweet* frequency and the number of followers to evaluate the quality of tweets [2][4], which are also used as major factors in our work. We show that the factors are highly important in evaluating the quality of tweets in this paper. Moreover, we consider the retweet step for original tweets, which is useful for ranking the tweets to be retrieved.

III. EXPERIMENTAL ANALYSIS OF RETWEET SERVICE

In our work, we have conducted an interesting experiment to analyze the characteristics of *retweet* service.



(a) The average number of users' followers

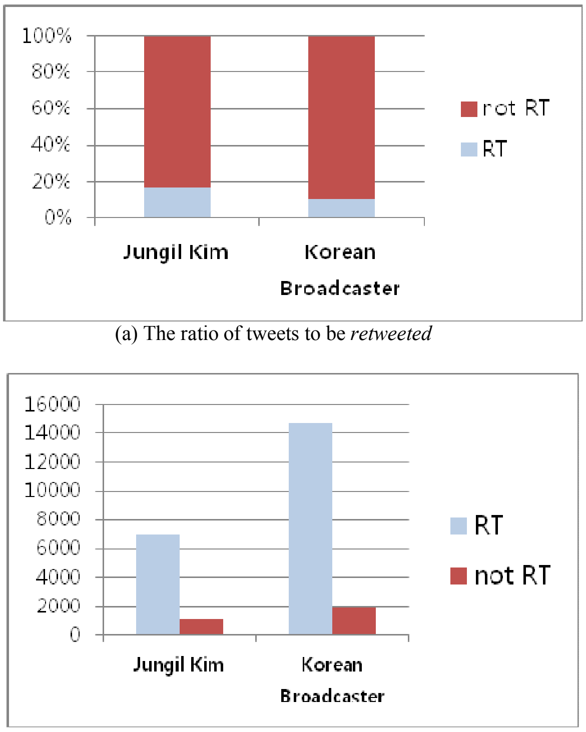
(b) The average number of posting articles

Figure 2 The average number of users' followers and posting articles

For the experiment, we have collected a large number of articles related to the current two hot issues: one is relevant to the death of Jungil Kim (who was a previous director of North Korea) with being collected for 36 hours from 2 PM, 17 Dec. 2011, and the other is relevant to an event of service shut down of a Korean major broadcaster with being collected for 36 hours from 9 AM, 17 Jan. 2012. People were much more interested in the former articles.

Figure 1 shows the frequency of *retweet* operations for tweets of the two hot issues. In this figure, *level-0* means the initial posting articles, and *level-i* means the *i*-th *retweeted* posting articles; that is, *level-i* means the frequency of *retweet* operations from *level-(i-1)*. As seen in the figure, the articles about 'the death of Jungil Kim' were *retweeted* at six steps, and the ones about 'the service shut down of a Korean broadcaster' at three steps. This implies that the articles related to hotter issues have more *retweet* operations.

Figure 2 shows the average number of a user's followers to perform *retweet* operations, and the average number of tweets posted by retweet users nearly at the *retweet* time. Figure 2(a) shows that both of the tweet sets are *not* obviously different in terms of the number of followers and the number of *retweet* steps. However, Figure 2(b) shows that the average number of tweets posted by users who have *retweeted* the articles related to "the service shut down of a



(a) The ratio of tweets to be *retweeted*

(b) The average number of users' followers for *retweeted* posting articles

Figure 3 The ratio of tweets and the average number of followers for *retweeted* posting articles

Korean broadcaster" is larger than the ones posted by users who have *retweeted* the articles related to "the death of Jungil Kim". This implies that the power users of Twitter are interested in various issues.

Figure 3(a) shows the ratio of the articles to be *retweeted* more than once to the articles not to be *retweeted*, in which RT and notRT mean the articles to be *retweeted* and not to be *retweeted*, respectively. In this figure, we have found that the 10~20% tweets among all the tweet articles are to be *retweeted*; of course, the ratio can change according to the degree of importance of articles. Figure 3(b) shows the average number of followers only for *retweeted* articles. As expected, we find that the average number of the followers of the user who has posted *retweeted* articles is larger than that of the user who has posted the articles not to be *retweeted*.

IV. THE SEARCH METHOD USING *RETWEET* INFORMATION

With considering the experimental patterns of Twitter data mentioned in Section III, we suggest the following assumptions, which are applied to our search function.

- The *retweet* operation is a way of dissemination for informative tweets worthy of sharing.
- As the number of a tweet user's followers increases, his/her articles are more likely to be *retweeted*. Consequently, the tweet user who has a lot of followers plays a great role in disseminating some valuable information.
- The tweets that have drawn great attention are more likely to be *retweeted*.

With the above assumptions, we propose a way of search method over tweets. Basically, searching tweets is to find out those having given query words. As in the Web search engines, it is necessary for Twitter search to rank the related articles in the order of some appropriate criteria such as their information values. As mentioned before, some tweet articles have very informative whereas others have very subjective or non-informative. Thus it is necessary to isolate highly informative articles by evaluating the frequency of *retweets*.

Let $\overline{D_q}$ be the set of articles that include the search keyword *q*. $\overline{D_q}$ includes not only the initial articles but also their *retweeted* articles; that is, let $\overline{D_q^0}$ and $\overline{D_q^i}$ be the initial articles and the *i*-th *retweeted* articles, respectively, then $\overline{D_q} = \bigcup_{i=0}^{\infty} \overline{D_q^i}$.

Under this assumption, a feasible way of evaluating the value of tweet articles is to the *retweet* frequency irrespective of the steps of *retweets*.

As a result, among the articles collected for *t* time, a set of the initial articles d_q^0 including the search keyword *q* can be given the following impact value.

$$impact(d_q^0, t) = \sum_{i=1}^{N(t)} \left| \overline{d_q^i} \right| \tag{1}$$

where $\overline{d_q^i}$ denotes a set of articles *retweeted* at the *i*-th step from d_q^0 , *N*(*t*) the number of *retweet* for *t* time and $\left|\overline{d_q^i}\right|$ the number of articles in $\overline{d_q^i}$. This idea has been applied to several search systems over Twitter data. For example, Korean twitter service such as *Joinmsn* help users to show the interested articles that has been *retweeted* the most frequently for some time [12-13].

The alternative way is to integrate the factor of *retweet* steps into Equation 1. Successive re-sending (i.e., *retweet*-ing) *retweeted* articles means that such articles are valuable so as to share with each other. This idea can be realized as the following equation by giving weight values at each *retweet* step.

$$impact(d_q^0, t) = \sum_{i=1}^{N(t)} \alpha_i \left| \overline{d_q^i} \right| \tag{2}$$

where α_i is the weight value at *i*-th *retweet* step. Therefore, this equation means that as α_i is given a larger value, the articles that have been successively *retweeted* can be assigned greater priorities for searching.

Furthermore, we cannot help considering the number of followers in evaluating the tweets. In general, some people who have a strong influence on others (i.e., have a large number of their followers) tend to be circumspect in *retweeting* tweet articles. Each of them is a sort of *information hub* in disseminating posting articles. Considering such an issue, we devise the following final equation for searching.

$$impact(d_q^0, t) = \sum_{i=1}^{N(t)} \sum_{d \in \overline{d_q^i}} \sum_{u \in pu(d)} \alpha_i \times \log_{10}(|\, followers(u)\,|) \tag{3}$$

where *pu*(*d*) denotes the user who has posted the tweet *d*, and *followers*(*u*) denotes the set of followers of the user *u*. In comparison with Equation 2, Equation 3 is considering the influencing power of a user who performs the *retweet* operation. Normally, there are wide variations in the number of followers. Thus, we use logarithm in Equation 3 to avoid undesirable fluctuation by too big (or small) values in computing the value of Equation 3

V. EMPRICAL RESULTS

We have performed a number of experiments in order to evaluate the proposed tweet search method. As for test data, we have prepared two kinds of tweets collections: one is a collection of more than 100,000 tweets about 'the death of Jungil Kim' and the other is a collection of more than 3,500 tweets about 'the service shut down of a Korean broadcaster'. With the test data, we have estimated the accuracy of ranking results by using Equations 1, 2, and 3. To evaluate the accuracy of ranking results, it is necessary to have correctly ranked lists for tweets. For this, we have made about 30 humans evaluate the quality of top 100 articles to be the most frequently retweeted.

As a measure of accuracy, we have used *nDCG* (Normalized Discounted Cumulative Gain) [14], which is commonly used to evaluate the effectiveness of Web search algorithms. Basically, this measure is to compare the current query results with perfect ranking results for different types of queries, which is defined for the query *q* as follows.

$$nDCG_q = M_q \sum_{i=1}^{K} (2^{rel(i)} - 1) / \log(1+i) \tag{4}$$

where *rel*(*i*) is the graded relevance of the query result at the position *i*, and it is evaluated with four discrete values, i.e., 0, 1, 2, 3. As the relatively upper part of query results are evaluated as higher values for *rel*(*i*), the *nDCG* values come to be higher. And, M_q is a normalized constant. Consequently, the *nDCG* value ranges from 0 and 1, and it has the value 1 when we obtain the best result.

Figure 4 shows the effect of the proposed methods in terms of *nDCG* measure, where *impact1*, *impact2*, and *impact3* denote the ranking methods by Equations 1, 2, and 3, respectively. In Equations 2 and 3, α_i which is the weight value at *i*-th *retweet* is set to 2^i as retweet step increases. As shown in the figure, the ranking method based on Equation 3 give the best results for both of the test collections. This means that the retweet frequency and the number of followers play an important role in evaluating tweets. Specifically, the *impact3* method outperforms the *impact1* method by 1.7% and 2.2% for the test collections about 'the death of Jungil Kim', and 'the service shut down of a Korean broadcaster', respectively. This is because the tweets related to the issue about 'the death of Jungil Kim' are mostly users' opinions rather than facts, and moreover some of them are the articles retweeted by followers.

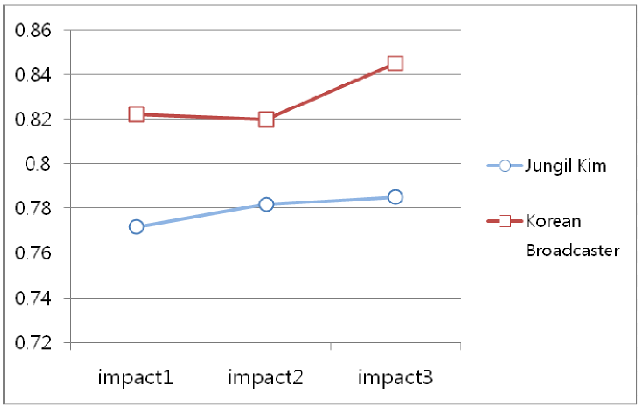


Figure 4 Accuracy of the ranking results in terms of nDCG

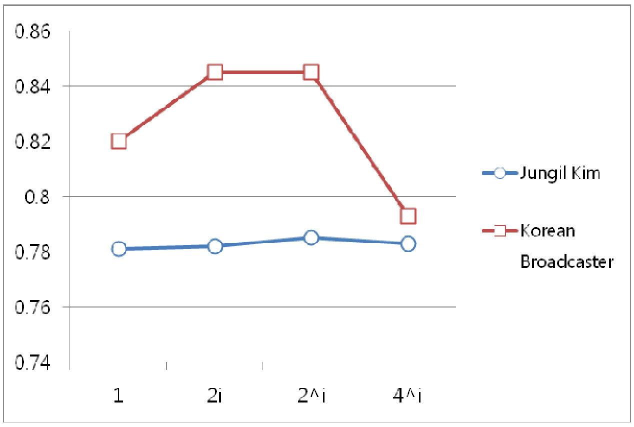


Figure 5 The changes of ranking accuracy from varying α_i

Figure 5 shows the changes of ranking accuracy from varying α_i when using Equation 3. The value of α_i is allowed to increase by 1, 2*i*, 2^i, 4^i according to the retweet step. As shown in this figure, the proposed method based on Equation 3 gives higher accuracy at the positions 2*i*, and 2^i over the test collection about 'the service shut down of a Korean broadcaster'. However, the method does not give any significant change from varying α_i over the test collection about 'the death of Jungil Kim'. This is also because most of the tweets related to the issue about 'the death of Jungil Kim' are users' opinions or retweeted articles. From the empirical results, we have found that retweet steps could not play a great role in evaluating the quality of tweets. We can expect that the proposed methods show relatively higher accuracy if the test collections contains more larger amount of informative news based on new ‘facts’.

VI. Summary and Future work

Recently, several studies on searching the Twitter data has been carried out as the social network services become very popular and influential. However, a lot of research issues need to be tackled since the Twitter data is extremely different from general Web documents.

In this paper, we proposed a novel method of searching Twitter data. The method considers two main factors: *retweet* and the number of users’ followers. Now, we plan to apply machine learning algorithms to the currently proposed method; this is because for more accurate search, it is necessary to automatically adjust the weight value α_i .

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (grant number: NRF-2011-0022445) and was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (grant number: NRF-2011-0025212) .

References

References

[1] TwitterEngineering, “200 million tweet per day”, http://blog.twitter.com/2011/06/200-million-tweets-per-day.html, [retrieved: Jan., 2012].

[2] R. Nagmoti and M. D. Cock, “Ranking Approach for Microblog Search”, Proc. of WI-IAT conference, pp. 153-157, 2010.

[3] A. Sarma, At. Sarma, S. Gollapudi, and R. Panigrahy, “Ranking Mechanisms in Twitter-like Forums”, Proc. of WSDM conference, pp. 21-30, Feb. 2010.

[4] H. W. Lauw, A. Ntoulas, and K. Kenthapadi, “Estimating the Quality of Postings in the Real-time Web”, Proc. of SSM conference, pp. 102-115, 2010.

[5] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search (2^{nd} Edition), ACM, 2011.

[6] http://lucene.apache.org/nutch/, [retrieved: Dec., 2011].

[7] J. Teevan, D. Ramage, and M. R. Morris, “#TwitterSearch: A Comparison of Microblog Search and Web Search”, Proc. of WSDM conference. pp. 35-44, 2011.

[8] M. Char, H. Haddadi, F. Benevenuto, and K. Gummadi, “Measuring User Influence in Twitter: The Million Follower Fallacy”, Proc. of International AAAI conference on Weblogs and Social Media, pp. 17-18, 2010.

[9] H. Kwak, C. Lee, H. Park, and S. Moon, “Finding Influentials Based on Temporal Order of Information adoption in Twitter”, Proc. of WWW conference, pp. 1137-1138. 2010.

[10] TunkRank, http://tunkrank.com, [retrieved: May, 2011].

[11] J. Weng and Q. He, “TwitterRank: Finding Topic-sensitive Influential Twitterers”, Proc. of WSDM conference, pp. 261-270, 2010.

[12] http://www.joinsmsn.com/, [retrieved: Jan., 2012].

[13] http://koreantweeters.com/, [retrieved: Jan., 2012].

[14] http://en.wikipedia.org/wiki/Discounted_cumulative_gain/, [retrieved: Mar., 2012].