Detecting Real-time Events using Tweets

Koichi Sato
Graduate School of Computer Science and Engineering
The University of AIZU
Aizu-Wakamatsu, Fukushima, 965-8580 Japan
Email: d8171103@u-aizu.ac.jp

Junbo Wang
Graduate School of Computer Science and Engineering
The University of AIZU
Aizu-Wakamatsu, Fukushima, 965-8580 Japan
Email: j-wang@u-aizu.ac.jp

Zixue Cheng
School of Computer Science and Engineering
The University of AIZU
Aizu-Wakamatsu, Fukushima, 965-8580 Japan
Email: z-cheng@u-aizu.ac.jp

***Abstract*—Big Data has been one of main topics in the field of computer science. Additionally, demand for observations of the real world in real time has increased to provide services or information to people accordingly. For example, when disaster occurs, government can appropriately respond to the disaster if the situations in the disaster-stricken areas are real-timely grasped. Although there are many kinds of blog services and they are functioning as one of Big Data source, Twitter is considered as the most active Big Data source. Users can feel free to post a tweet anywhere in real time, since twitter limits a tweet to 140 characters. In this paper, a scheme is proposed which can detect what happens in real world in real time only by analyzing tweets as Big Data and let a user know the event. To this end, the following problems has to be solved. They are a) quantifying importance of words accurately and b) evaluating the quantified values dynamically. As the solutions for the problems, two new methods are proposed which are the Extended Hybrid TF-IDF and the Remarkable Word Detecting Method, and they are used in the proposed scheme. Finally an experiment is executed to evaluate the proposed methods and scheme.**

I. Introduction

Recently, the demand for observations of what happens in the real world through big data in real time is increasing. For example, when disaster occurs, it is more important to grasp the situations in the disaster-stricken areas quickly. Real-time situation detections can lead lifesaving, effective reconstruction of emergency base stations for mobile phones, and effectively providing emergency relief goods. Sensor networks may be a choice to satisfy the demand if a lot of sensors can be deployed all over the world. However this solution is restricted by the enormous costs. A micro-blog, i.e., twitter, as a new type of "social sensing" has attracted a lot of researchers to understand real world situation. For example, Inouye et al. researched twitter summarization algorithms, and proposed the Hybrid TF-IDF [1]. TF-IDF is designed to quantify importance of a word included in a document and good at sentence analysis, but it doesn't fit for tweet analysis due to the limited number of characters in a tweet. Twitter limits the number of characters in a tweet to 140 characters. It is not a good feature for the TF-IDF, because it is difficult to get a term frequency of a word from a very short sentence. The Hybrid TF-IDF is a kind of TF-IDF improved to solve the problem. The Hybrid TF-IDF can quantify importance of a word included in a tweet, but is not so accurate. Takeshi et al. proposed the real-time event detection system using tweets [2]. It can detect an event but needs classifying a tweet into a positive class and a negative class by using a support vector machine in the beginning of the process. Positive and negative examples as a training set have to be prepared and fed into the support vector machine requires to do the classification correctly. Rui et al. proposed a Twitter Based Event Detection and Analysis System(TEDAS)[3]. It can detect events and places where the events occur. However it requires that sets of keywords are prepared in advance to detect them. Therefore it can only detect events prepared sets relate to. Different with the above approaches, this paper proposes a scheme to let a user know an event which happen in real world in real time by extracting important words from gathered tweets. Note that a remarkable word is defined as a word which expresses or relates to the event in this paper. To this end, the following two problems are mainly targeted in this study:

- How to quantify importance of a word included in a tweet
- How to evaluate the quantified importance efficiently to detect a remarkable word

To solve the first problem, in this paper we propose an Extended Hybrid TF-IDF. Although it is developed being inspired by the Hybrid TF-IDF, it can be more accurate than the original method. Using the Extended Hybrid TF-IDF, the proposed scheme can quantify importance of words included in tweets. To solve the second problem, we further propose the Remarkable Word Detecting Method. Efficiently to detect a remarkable word, it can dynamically provide a threshold to evaluate the quantified importance. The threshold is generated from moving averages of Extended Hybrid TF-IDF values.

II. Outline

The proposed scheme is mainly divided into three parts which are Tweet Preprocessing (TP), Importance Degree Calculating (IDC) and Remarkable Words Detecting (RMD) as shown in Fig.1. TP is designed for extracting information from tweets. It uses Twitter Streaming API and the Morphological Analysis Engine for the purpose. Its works are as follows:

- It takes tweets with Twitter Streaming API.
- It extracts nouns from the tweets with the Morphological Analysis Engine.

978-1-5090-4240-1/16/$31.00 ©2016 IEEE

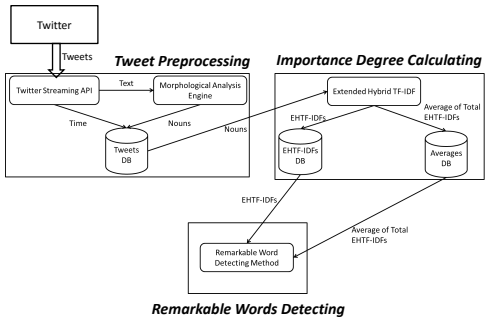

Fig. 1. Outline

- It stores the extracted nouns into Tweet DB.

The second part is IDC, which is designed for getting Extended Hybrid TF-IDF values of the nouns and an average of the values. It uses Extended Hybrid TF-IDF to get Extended Hybrid TF-IDF values. The detailed explanation will be discussed in section III. Its works are as follows:

- It gets Extended Hybrid TF-IDF values of the nouns with Extended Hybrid TF-IDF.
- It stores the values into TF-IDF DB.
- It calculates an average of the values.
- It stores the average into Average DB.

RMD is designed for detecting remarkable words based on currently gotten Extended Hybrid TF-IDF values and some averages of currently and previously gotten them. It works as follows:

- It gets moving averages of each value of Extended Hybrid TF-IDF.
- It gets a moving average of the averages, which can be a threshold of a remarkable word.
- It considers words, such that moving averages of the words exceed the threshold, as remarkable.

III. Proposed Methods

This paper proposes two methods which are the Extended Hybrid TF-IDF and the Remarkable Word Detecting Method. First one is proposed to quantify importance of a word in numerical form more accurately than Hybrid TF-IDF. Second one is proposed efficiently to detect remarkable words.

A. Extended Hybrid TF-IDF

The Extended Hybrid TF-IDF, shorted as EHTF-IDF, is an algorithm to quantify importance of words included in a tweet. The algorithm is constructed by two calculations, which are Extended Hybrid TF and Extended Hybrid IDF. They are shorted as EHTF and EHIDF respectively. An EHTF value is gotten based on appearance frequency of a word among current tweets. If the value is high, the word can be considered as an important word. On the other hand, an EHIDF value is gotten based on generality of a word among tweets. The tweets include not only current tweets but also past tweets. In this calculation, if the tweets include many numbers of one word, an EHIDF value of the word is low. It means that a generally appearing word among the tweets is not important, since a word of this kind is a generally used word. According to EQ.1, EHTF-IDF quantifies importance of a word multiplying an EHTF value of the word by an EHIDF value of the word. This multiplying is calculation to give high importance to a word which frequently appears in a current tweets but is not a general used word.

$$EHTF\text{-}IDF = EHTF * EHIDF \tag{1}$$

1) Extended Hybrid TF: How to get an EHTF value of a word is the same with Hybrid TF-IDF's. First all current tweets are organized into a document. Second count a total number of appearance of all words in the document. Third count a number of appearance of a word in the document. Finally an EHTF value of the word is gotten by dividing the number of appearance by the total number. This calculation is as the following equation:

$$EHTF = \frac{NAW}{TNAAW} \tag{2}$$

Where NAW is a number of appearance of a word in a document, and TNAAW is a total number of appearance of all words in a document.

2) Extended Hybrid IDF: How to get an EHIDF value of a word is different from Hybrid TF-IDF's. In the proposed algorithm, all current tweets are organized into a document as with the EHTF calculation, and then each past tweet is respectively considered as past documents. The past documents are used to evaluate generality of a word which included in the current document according to EQ.3. On the other hand, in Hybrid TF-IDF's, each current tweet is respectively considered as current documents and past tweets are not used. The Hybrid IDF calculation uses these documents to evaluate generality of a word which included in tham. In this way, it is difficult that this processing works effectively. It is the reason why the proposed method is more accurate than Hybrid TF-IDF.

$$EHIDF = \log \frac{TNPD}{NPD} + 1 \tag{3}$$

where TNPD is a total number of past documents, and NPD is a number of past documents including a word which is included in the current document.

B. Remarkable Word Detecting Method

Degrees of importance of words can be represented in numerical form by EHTF-IDF. As it is no more than numerical values, it is impossible to detect remarkable words only from the numerical values. Therefore threshold is required for the detection. The Remarkable Word Detecting Method, shorted as RWDM, is designed for dynamically getting the threshold and then detecting remarkable words efficiently. This method uses two moving averages, which are EHTF-IDF Value Moving Average and Total EHTF-IDF Values Moving Average, respectively shorted as EVMA and TEVMA. The reason why the proposed method uses the moving averages is to reduce a

rate of false detection. Sometime, an EHTF-IDF value includes noise. As a moving average is less affected by noise, this method uses EVMA and TEVMA. In this method, TEVMA is used as threshold, so a word can be judged important if EVMA of the word exceeds TEVMA.

Require:

TEA is a set of averages of total EHTF-IDF values, and EV is a set of EHTF-IDF values of a word. x is a number of averages of total EHTF-IDF values used to get TEVMA. y is a number of EHTF-IDF values of a word used to get EVMA.

Ensure:

$TEVMA = 0$
for $i = 0$ to $x - 1$ **do**
  $TEVMA = TEVMA + TEA_i$
end for
$TEVMA = TEVMA/x$
$EVMA = 0$
for $j = 0$ to $y - 1$ **do**
  $EVMA = EVMA + EV_j$
end for
$EVMA = EVMA/j$
if $EVMA > TEVMA$ **then**
  This word is considered as a remarkable word
end if

C. Example

To explain this method clearly, we give an example by assuming that:

- EHTF-IDF calculation was serially executed twelve times by the minute
- EHTF-IDF got EHTF-IDF values of words by each calculation
- An average of the total values was gotten by after each calculation
- EVMAs are gotten from the newest value of a word and the two near past values of a word.
- TEVMAs are gotten from the newest average and the two near past averages.
- A great earthquake occurs in Sendai eight minutes later from the beginning.

According to the assumptions, there are the twelve EHTF-IDF values by a word, the twelve averages, the ten EVMAs by a word and the ten TEVMAs. Table I includes the results of the twelve averages and the twelve EHTF-IDF values of three words which are Sendai, Earthquake and Comic. Sendai is one of a cities in Japan. Fig. 2 is made from Table I. Table II includes the TEVMAs and EVMAs of the words. Fig. 3 is made from Table II. First of all, look at Fig. 2. There are three lines which are a Sendai line, a Earthquake line, a Comic line and an average line. The Sendai line and the Earthquake line are always low and under the average line before nine. It is natural, since Sendai is no more than one of Japanese local cities as mentioned above, and a word of Earthquake does not appear in tweets many time if earthquake does not occur. These values increase dramatically at nine, and their lines exceeds the average line. After that, these lines keep above the average line. The reason is the sixth assumption, so it leads the movement. On the other hand, a Comic line basically keep under the average line throughout the almost all of time, but the line exceeds the average line only once. It is noise, since the EHTF-IDF value of Comic goes back to an usual range of the EHTF-IDF value immediately. An EHTF-IDF value is basically affected by an event which occurs in the real world. If the increasing is leaded to by the event and it is not noise, the comic line has to keep above the average line for a while at least, since the event is not lost the attention of the public immediately. If the average line is used as the threshold, the judgment is greatly affected by noise like this one. Next, look at Fig. 3. The Sendai line and the Earthquake line exceed the TEVMA line, and they are correctly judged important. On the other hand, the Comic line does not exceed the TEVMA line, and it is judged not important. It is the reason why TEVMA is used as the threshold.

TABLE I
EHTF-IDF VALUES OF FOUR WORDS AND AN AVERAGE

Time	Sendai	Earthquake	Comic	Average
1	1.557262848	1.423669487	2.960239587	4.14514147
2	2.197407972	0.424552981	2.983769721	4.226861877
3	0.520374003	1.479191256	3.034343401	4.065876742
4	1.730000062	2.241758709	2.512462758	4.080799658
5	0.072371223	0.712560369	4.8	4.143385595
6	1.330507504	0.019836056	2.823316796	4.46446488
7	0.341714873	1.925201166	2.559435068	4.129062209
8	1.831404064	2.131805909	2.511471342	4.348951954
9	4.881125632	4.721226219	2.75245761	4.12504995
10	4.96148779	4.816891669	3.319433913	4.263936493
11	4.6090735	4.61154061	3.281607591	4.317679894
12	4.657209485	4.994994583	3.178982052	4.454033985

TABLE II
MOVING AVERAGES OF THE FOUR VALUES

Sendai	Earthquake	Comic	Average
1.425014941	1.109137908	2.992784236	4.14596003
1.482594012	1.381834315	2.843525293	4.124512759
0.774248429	1.477836778	3.448935386	4.096687332
1.04429293	0.991385045	3.378593185	4.229550045
0.5815312	0.885865863	3.394250621	4.245637562
1.16787548	1.35894771	2.631407736	4.314159681
2.351414856	2.926077765	2.607788007	4.201021371
3.891339162	3.889974599	2.861120955	4.245979466
4.817228974	4.716552833	3.117833038	4.235555446
4.742590258	4.807808954	3.260007852	4.345216791

TABLE III
THE FOUR PARAMETERS

First Parameter	1 min.
Second Parameter	360 tables
Third Parameter	5 TF-IDF values per a word
Fourth Parameter	5 averages

IV. IMPLEMENTATION UNDER THE EVALUATION

In this section, it is showed as an example how to build Real-time Event Detecting System from Japanese Tweets

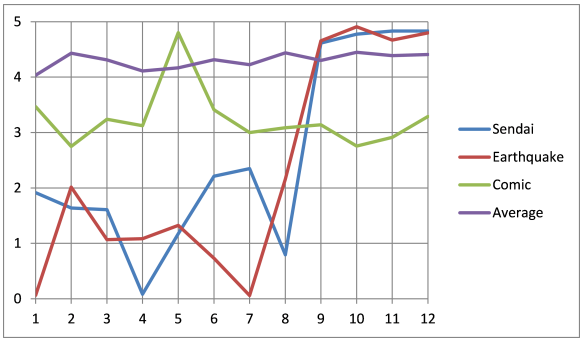

Fig. 2. The EHTF-IDF values

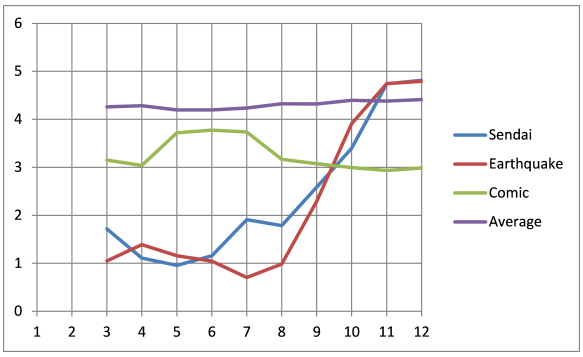

Fig. 3. Moving Averages of the values

according to proposed scheme. Each part mentioned in the section II is implemented as each module. This system is constructed by the three modules, which are Tweet Preprocessing Module, Importance Degree Calculating Module and Remarkable Words Detecting Module, and needs four arbitrarily allocated parameters which are hereinafter referred to as Par.1, Par 2., Par.3 and Par.4 respectively. Explanation of the parts and parameters are provided step by step in this section. Their values are referred to Table III.

A. Tweet Preprocessing Module

The Tweet Preprocessing module is continuously invoked to take a tweet using Twitter Streaming API, and then extract nouns from the tweet with the Morphological Analysis Engine. The reason of the extracting is that a noun has the highest possibility of relating to an event being occurring in the real world. The extracted nouns are stored as a record on a table within the Tweets DB. The table is constantly renewed new one at an interval which depends on Par.1. Par.1 is the interval to renew the table.

B. Importance Degree Calculating Module

Once a table is renewed, the Importance Degree Calculating module is invoked, and acquires current and previous tweets. The current tweets are on the newest table, which are analysis subjects. The previous tweets are on some of the previous tables. It depends on Par.2 how many previous table, in reverse chronological order, provide their tweets for the module. Par.2 is a number of previous tables which provide their tweets for this module. The previous tweets are used to get an EHIDF value in the Extended Hybrid TF-IDF. A reason why it is invoked and provided these tweets includes;

- To get EHTF-IDF values of all words included in current tweets with the Extended Hybrid TF-IDF and the provided tweets
- To get an average of the values
- To store the values on TF-IDF values DB
- To store the average on Average DB

C. Remarkable Words Detecting Module

The Remarkable Words Detecting module is invoked after the Importance Degree Calculating module. It acquires some of the EHTF-IDF values per word and some of the averages in reverse chronological order. Par.3 is a number of the provided EHTF-IDF values per word. Par.4 is a number of the provided averages. This module executes Remarkable Words Detecting Method for each word. Remarkable Words Detecting Method detects remarkable words efficiently with the provided values and averages. Finally the detected remarkable words are notified.

V. Evaluation

In this section, a simple experiment is executed to evaluate the capability of the proposed scheme, the Extended Hybrid TF-IDF and the Remarkable Word Detecting Method. First of all, four Real-time Event Detecting Systems are prepared. Their configurations are as follows:

- Extended Hybrid TF-IDF and Remarkable Word Detecting Method(called ER)
- Extended Hybrid TF-IDF and the Average(called EA)
- Hybrid TF-IDF and Remarkable Word Detecting Method(called HR)
- Hybrid TF-IDF and the Average(called HA)

The first one is the same with the Real-time Event detecting System proposed at previous section. The second and fourth one simply use the average of Extended Hybrid TF-IDF or Hybrid TF-IDF values as a threshold.

A. Experiment

A whole picture of this experiment is as follows:

- The Real-time Event detecting Systems are used for this experiment.
- The 2015 Kohaku Utagassen, which is a popular song program on Japanese television, is the analysis subject.
- Their detections are success if a name of a singer is detected as a remarkable word while he or she is singing.
- Finally the event detecting capabilities are gotten by dividing a number of the successes by the total number of the singers who perform on the program.

This experiment uses the Real-time Event Detecting Systems. The Kohaku Utagassen is the most famous popular song program in Japan. It is broadcasted on the last day of the year every year. The viewing rating of the program in 2015 was 39.2 percent. As it can be considered as an event of national concern in Japan, it is used as the analysis subject. While the

program has been broadcasted, they have detected remarkable words. if a name of a singer is detected as a remarkable word while he or she is singing, their detections are considered successful.

B. Extended Hybrid TF-IDF

First the Extended Hybrid TF-IDF is evaluated. The contribution of the Extended Hybrid TF-IDF is to be able to quantify importance of a word included in a tweet more accurately than the Hybrid TF-IDF. Therefore the accuracy of the Extended Hybrid TF-IDF is compared with the accuracy of the Hybrid TF-IDF for this evaluation. The accuracy rate with the Hybarid TF-IDF is gotten by the following equation.

$$The\ accuracy\ rate = \frac{ER + EA}{HR + HA} \tag{4}$$

According EQ.4 and Table IV, The accuracy rate with Hybrid TF-IDF can be gotten by the following calculation.

$$The\ accuracy\ rate = \frac{84.6 + 84.6}{67 + 80.8} \approx 114.5\% \tag{5}$$

From the result, it is found that the Extended Hybrid TF-IDF can quantify importance of a word included in a tweet approximately 1.145 times more accurately than Hybrid TF-IDF.

C. Remarkable Word Detecting Method

Second the Remarkable Word Detecting Method is evaluated. The contribution of the Remarkable Word Detecting Method is to be able to provide the threshold with higher noise tolerance than a way of using an average of the Extended Hybrid TF-IDF or the Hybrid TF-IDF values as the threshold. However this evaluation is very difficult, because getting F-measures of them is virtually impossible due to the lack of ground truth. Therefore it is impossible to evaluate them by a comparison of their F-measures. For example, Although Kumar, Liu, Mehta and Subramaniam researched identifying events by using Twitter Streaming API[4], they could not evaluate it precisely due to the lack. In their case, they only verified whether it could detect the top stories of the day. Similarly it is difficult to evaluate the proposed method completely precisely. However, in our experiment, it is considered as a better feature that a name of singer is detected in a fewer number of remarkable words. Table V shows numbers of detected remarkable words which only include remarkable words of the singers who are detected by all of systems. The efficiency rate with the way of using the average is gotten by the following equation.

$$The\ efficiency\ rate = \frac{EA + HA}{ER + HR} \tag{6}$$

According EQ.6 and Table V, The efficiency rate with the way can be gotten by the following calculation.

$$The\ efficiency\ rate = \frac{838 + 1399}{134 + 618} \approx 274\% \tag{7}$$

TABLE IV
THE EVENT DETECTING RATES OF THE FOUR SYSTEMS

Systems	Events detecting Rates
ER	84.6%
EA	84.6%
HR	67%
HA	80.8%

TABLE V
NUMBERS OF DETECTED REMARKABLE WORDS

Systems	Numbers of detected Remarkable words
ER	134
EA	838
HR	681
HA	1399

From the result, it is found that the Remarkable Words Detecting Method is approximately 2.74 times more efficient than the way of using the average.

D. Proposed Scheme

Finally the proposed scheme is evaluated. The contribution of the proposed scheme is to be able to detect a remarkable word as an event effectively. Therefore it can be simply evaluated by getting an event detecting rate. The event detecting rate is gotten by dividing a total number of the successes by a total number of the singers who perform on the program. The number of the singers was fifty-two, so the rate is gotten by the following equation.

$$The\ event\ detecting\ rate = \frac{TNoS}{52} \tag{8}$$

where TNoS is a total number of the successes. From Table IV, it is found that ER and EA have the highest detecting results. Therefore the Extended Hybrid TF-IDF is more accurate than Hybrid TF-IDF. In terms of evaluating the Remarkable Word Detecting Method, the result of ER seems the same with the result of EA. However, from Table V, ER is approximately 6.25 times more efficient than EA.

VI. CONCLUSION

In this paper, a scheme is proposed which can detect what happens in real world in real time only by analyzing tweets. To this end, there are two problems which are how to quantify importance of a word in a tweet accurately and evaluate the quantified importance efficiently to detect a remarkable word. As the solutions for the problem, the Extended Hybrid TF-IDF and the Remarkable Word Detecting Method are proposed. The Extended Hybrid TF-IDF can quantify importance of a word included in a tweet approximately 1.145 times more accurately than Hybrid TF-IDF. The Remarkable Words Detecting Method is approximately 2.74 times more efficient than the way of simply using an average of Extended Hybrid TF-IDF values. In the experiment, the Real-time Event Detecting Systems which is build according to the propose scheme can detect 84.6% of events.

ACKNOWLEDGMENT

This research was partially supported by JST, Strategic International Collaborative Research Program, SICORP, entitled Dynamic Evolution of Smartphone-Based Emergency Communications Network, from 2015 to 2018.

REFERENCES

[1] D. Inouye and J. K. Kalita, *"Comparing twitter summarization algorithms for multiple post summaries"*, in 2011 IEEE International Conference on Privacy, Security, Risk and Trust, and IEEE International Conference on Social Computing. IEEE, 2011, pp. 298-306.

[2] T. Sakaki, M. Okazaki, and Y. Matsuo, *Earthquake shakes twitter users: Real-time event detection by social sensors,*, in Proc. 19th International Conference on WWW2010, pp. 851860, 2010.

[3] Rui LI, Kin Hou Lei, Ravi Khadiwala, Kevin Chen-Chuan Chang, *TEDAS: a Twitter Based Event Detection and Analysis System,*, in 2012 IEEE 28th International Conference on Data Engineering, pp. 1273 - 1276, 2012.

[4] Shamanth Kumar, Huan Liu, Sameep Mehta and L. Venkata Subramaniam, *Exploring a Scalable Solution to Identifying Events in Noisy Twitter Streams*, in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15, pp. 496 - 499, 2015.