Sensing Microblog for Effective Information Extractions

Sindur Patel[1], Nirav Bhatt[1], Chandni Shah[1], Rutvika Nanecha [2]

[1] Department of Information Technology , Charotar University of Science & Technology, Changa, Gujarat India
[2]Department of Information Technology , Charotar University of Science & Technology, Changa, Gujarat India
sindurpatel@gmail.com ,niravbhatt.it@charusat.ac.in, chandnishah.it@charusat.ac.in, rutvi1710@gmail.com

Abstract. The SMERP 2017 data challenge track given a set of tweets posted during Italy earthquake. For retrieving more relevance information respect to user interest profile in this paper provide BM25 and word2vec techniques for retrieving relevance information from twitter stream. This techniques aim is to find real-world and most relevance information respect to the query. For retrieving most relevant information used query expansion techniques. Information rank retrieval techniques BM25 find important data and give the final score to that information with respect to user interest profile. The result of our method in this task shows this is an effective method.

Keywords: Real-time data, relevance information, microblog, twitter stream.

1 Introduction

Microblog is a broadcast medium that allows the user to post short and frequent message [5]. It's a communication way compared with traditional information, microblogging has gained increased attention among people, organization, research scholars in distinct disciplines.

Twitter is currently fast growing micro-blogging services, with more than 140 or 150 million users producing over 400 or 500 million tweets per day [5]. It is an unable to twitter user for update status or tweets, no more than 140 characters to networks of follower using various communication services. Tweets size are limited, Twitter is updated millions of time a day by twitter user all over the world[5], and its data varies hugely based on user interest and behaviors. So twitter data have huge amounts of information scaling from news, events etc.

Twitter Provides timely or real information of any event. Observing, keeping and analyzing this content of user-generated data can yield new unprecedented important information, which not available from traditional media [5]. Tweets do the live reporting of any event [6] means finding the information what people are talking away from some conferences, debates, sporting events etc.

2 Challenges

A major problem of twitter is no any rules to post tweets, information's or status so some people provide false, incorrect information about some events. Many numbers of spellings, grammar error, and the use of not a proper sentence structure and mixed language so people can't distinguish important data from unused data. Not all tweets are relevant to the user query or interest profile.

One-way communication. Twitter often acts as a one-way communication platform. Twitter used by celebrities, TV shows, companies and websites to simply get the word out. It is not used for relationship building.

3 Information Extraction System

In this section introduce system architecture for retrieve tweets and do the scoring of tweets based on the query. The system contains four components [2].

3.1 Feature Extraction Components

It extracts a feature from twitter respect to TREC-API (Stream API and Rest API). After obtaining twitter streams we apply preprocessing and filtering to reduce tweets we need to process.

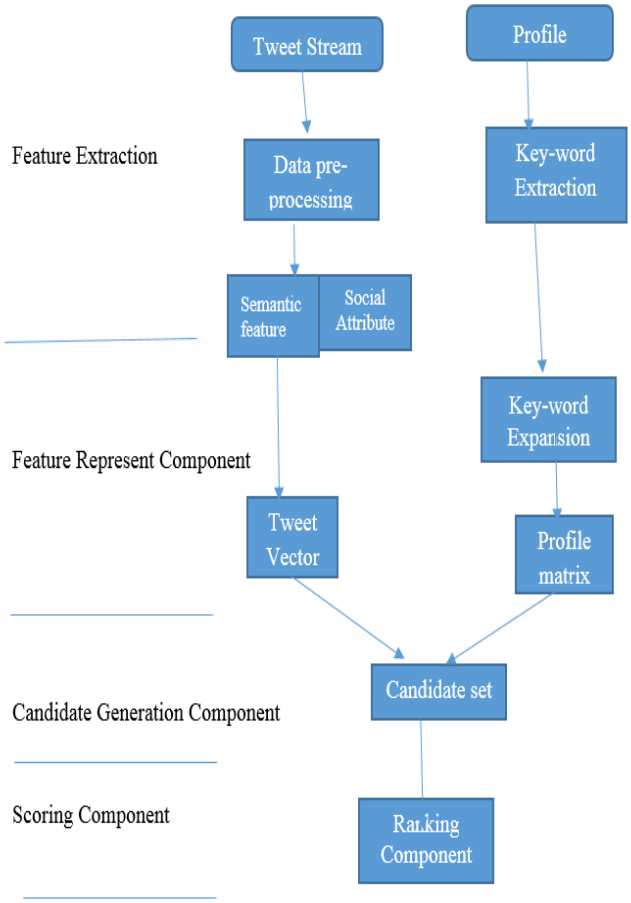

Fig. 1. Different System Components

3.2 Feature Representation Components

It represents and expands semantic feature by different expansion techniques. After extracting tweet we need to represent those features in a format so it is suitable to calculate relevance score between tweet and profile.

3.3 Candidate Generation Components

We classify tweet into the most relevance profile or remove it directly if it does not match any profile.

3.4 Scoring and Pushing Components

By the semantic feature (consider only verbs and nouns in tweet text) and social media attributes we got score semantic (C_i) and quality (Q_i) so final score $S_i = C_iQ_i$.

4 Query Expansion Entries

The query provided by the user is not in a structured and that is incomplete. So then we need to expand that query and do the correct for the better relevance information.

The main problem in retrieval is that query is short and unable to accurately describe user's information needs. So the solution to this problem is query Expansion [3], [4].

4.1 Word2Vec

For retrieving better result we have used word2vec model. Word2vec model used to produce word embeddings [8]. Predict surrounding words of all word or every word. This model use document or data to train a model maximizing conditional probability of context given the word. Take an input as a large data of text and produce a vector space. So we have expanded the query using this model and then after finding the result.

5 Query Relevance Model

The query provided by the user is not in a structured and that is incomplete. So then we need to expand that query and do the correct for the better relevance information.

5.1 BM25

BM25 is the best matching bag of word retrieval ranking function[6] that ranks an information based on the user interest profile or query words appearing in each document's information[1,2]. Developed in the Okapi system in London University. BM25 formula contains many parameters which need to be tuned from relevance assessment [9].Given a user interest profile P, containing keywords p_1, … , p_n the BM25 score of document D is

$$\text{Score (D, P)} = \sum_1^n IDF(p_i) \cdot \frac{f(p_i,D)(k_1+1)}{f(p_1,D)+k_1(1-b+b.\frac{|D|}{avgd\ 1})} \quad (1)$$

Where $f(P_i, D)$ is p_i's term frequency in document D, |D| is the length of document D in words, and avgdl is the average document length in the text collection from which documents are drawn[7]. k1 and b are default parameters, usually chosen, in absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and $b \in [0.5, 0.8]$[7]. In our case, we have used $k_1 = 1.2$ and $b = 0.5$. IDF (q_i) is the IDF weight of the query term q_i[7].

6 System Evaluation Result

Our result has been evaluated by the SMERP 2017 data challenge track. The evaluation score in terms of bpref, precision@20, Recall@1000 and MAP has been given by the SMERP as 0.2021, 0.1625, 0.1830 and 0.0180 respectively. The evaluation scores of the system without query expansion have been reported as 0.0218, 0.0875, 0.0218 and 0.0072 respectively. Below table shows the result. Our run_id is charusat_smerp17_1.

Table 1. SMERP Level-1 Evaluation Result Table

SL No	Team-id	Run-id	Run type	bpref	Precision@20	Recall @1000	MAP
1	DCU	dcu_ADAPT_run3	Semi-automatic	0.4407	0.1750	0.1256	0.0338
2	USI	USI_1	Semi-automatic	0.3286	0.5375	0.3183	0.1403
3	DAIICT	daiict_irlab_2	Semi-automatic	0.3171	0.2250	0.3171	0.0417
4	RU	rel_ru_nl_lang_analy	Semi-automatic	0.3153	0.2125	0.1913	0.0678
5	DAIICT	daiict_irlab_1	Semi-automatic	0.3074	0.2125	0.3015	0.0391
6	CSPIT	charusat_smerp17_1	Semi-automatic	0.2021	0.1625	0.1830	0.0180

7 Conclusion

In this paper present the research in the area of information retrieval on the microblog. We have worked on Italy earthquake data that given by SMERP 2017 data challenge track. We have submitted two runs, without word2vec and using word2vec. So we observed that using query expansion technique word2vec showed a better result. Train word2vec using large data and find the improvement in the result.

References

1. Zhu, X., Huang, j., Zhu, S., et al.: NUDTSNA at TREC 2015 Microblog Track: A Live Retrieval System Framework for Social Network based on Semantic Expansion and Quality Model. In: TREC (2015)
2. Bagdouri, M., W. Oard, D.: CLIP at TREC 2015: Microblog and LiveQA. In: TREC (2015)
3. Qiang, R., Fan, F., Lv, C., Yang, J.: Knowledge-based Query Expansion in Real-Time Microblog Search. arXiv preprint arXiv:1503.03961 (2015)
4. Lau, C.H., Li, Y., Tjondronegoro, D.: Microblog retrieval using topical features & query expansion. In: TREC (2011)
5. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. Computational Intelligence, 31(1), 132-164 (2015)
6. Why Twitter ?, http://webtrends.about.com/od/twitter/a/why_twitter_uses_for_twitter.htm
7. Okpi BM25, https://en.wikipedia.org/wiki/Okapi_BM25
8. Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
9. Svore, K. M., & Burges, C. J.: A machine learning approach for improved BM25 retrieval. In: Proc. 18th ACM conference on Information and knowledge management, pp. 1811-1814. ACM (2009)