Information Extraction from Microblogs

Prashant Bhardwaj
Computer Science and Engineering
National Institute of Technology Agartala
cse.pbh@gmail.com

Partha Pakray
Computer Science and Engineering
National Institute of Technology Mizoram
parthapakray@gmail.com

ABSTRACT

The micro blogging sites contain the emotion and expression of the public in raw format. The data can be used to extract much meaningful information that could be used to develop technologies for future use. There are numerous micro blogging sites available these days that are used in different contexts. Some are used basically for conversation, some for image and video sharing, and some for formal and official purposes. Twitter is one of the most outspoken platform for sharing the emotions and comments on almost every topic starting from sport to entertainment, religion to politics and many more. The paper attempts to extract information from the database of the tweets collected from Twitter. The task is to develop methodologies for extracting tweets that are relevant to each topic with high precision. This paper presents the nita_nitmz team participation in FIRE 2016 Microblog track.

CCS Concepts

- **Computing methodologies ~ Natural language Processing**
- **Information systems ~ Information extraction**

Keywords

Information Retrieval ; Micro blogging ; Twitter

1 INTRODUCTION

The faster growth of Internet in current period provides new sources of information. Now a day's people prefer to express themselves more often on social sites than any print media. The idea of Information Extraction from Microblogs Posted during Disasters was introduced by Sarah Vieweg et.al., in 2010 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems [1]. Henceforth it has become one of the most researched topics considering the possibilities it contained for the proper accessing of any incidence. The importance of the said topic can be attributed to the fact that people rarely provide any false information in the social sites and pour their emotions according to their knowledge and wisdom. This paper presents the experiments carried out at National Institute of Technology Agartala as part of the participation in the Forum for Information Retrieval Evaluation (FIRE) 2016 in Information Extraction from Microblogs Posted during Disasters [14]. The experiments carried out by us for FIRE 2016 are based on stemming, zonal indexing, theme identification, TF-IDF based ranking model and positional information. The data contained 48845 tweets out of 50,000 tweets mentioned in the workshop website. Query was provided by the organizing committee and each query was specified using title, narration and description format.

2 RELATED WORKS

The problem of Information Extraction from Microblogs Posted during Disasters is researched for a couple of years starting from 2010 by Sarah Vieweg et.al. [1] and Leysia Palen et.al.[2]. But there has been tremendous work since then and a new field of information retrieval has come into existence. Sudha Verma et.al. wrote on Situational Awareness through tweets [3]. The research on location of disaster hit area, the response and the information extraction has been going on since then [4][5][6][7]. One of the important part of the information retrieval part is the part of speech tagging in the code mixed microblog data[8][9][10][11][12][13]. Several researchers even work on information extraction from mixed script analysis in social media websites and forums. English used to dominate the micro blogging sites previously such as Twitter and Facebook.

3 TASK DESCRIPTION

A large set of microblogs (tweets) posted during a recent disaster event was be made available, along with a set of topics (in TREC format). Each 'topic' identified a broad information need during a disaster, such as – what resources are needed by the population in the disaster affected area, what resources are available, what resources are required / available in which geographical region, and so on. Specifically, each topic contained a title, a brief description, and a more detailed narrative on what type of tweets will be considered relevant to the topic. The participants are required to develop methodologies for extracting tweets that are relevant to each topic with high precision as well as high recall.

The data contained:

- Around 50,000 microblogs (tweets) from Twitter, those were posted during the Nepal earthquake in April 2015. Only the tweetids of the tweets was provided, along with a script that was be used to download the tweets using the Twitter API. Out of 50,000 tweets only 48845 could be downloaded on my experimental setup.
- A set of 5 – 8 topics in TREC format, each containing a title, a brief description, and a more detailed narrative.

4 METHODOLOGY

For the given task we created the required searching configuration on Apache Nutch 0.9 which is a highly extensible and scalable open source web crawler software project. The implementation of the said task is done in two steps. First is creating the searching environment, Secondly apply appropriate test queries to search the results from the previously configured Nutch using Tomcat server.

4.1 Preparation of the data

The python script provided by the organizers helped to get the tweets. But file generated was of *json* type. So another script had to be developed to extract the tweets from the json file. The json file contained the tweets, tweet_id and many more metadata. First problem arises when after extracting tweets from the json file, we found out that only 48845 tweets were downloaded via the given script. As per the norms of the Apache Nutch, the tweets had to separated into different files. Since the task was to extract relevant tweets , the files containing the tweets had to be named according to their tweet_ids. We developed two files one containing the tweets only and other containing the tweet_ids. After that we wrote a code to take the tweet_id from one file , create a text file with that name , take a tweet from another file and store it in the newly created file. But due to some unavoidable errors, only 48815 file could be created, each having the name as tweet_id and containing the corresponding tweet inside.

4.2 Crawling using Nutch

We used another code to store the addresses of the different file in the urls.txt file. We started the crawling part using the nutch. It works in following steps:

Injecter: All the URLs are taken by the injector from the seed.txt (here urls.txt) file, compares urls with regex-urlfiler regex and update crawldb with supported urls.. The crawldb maintains information on all known URLs (fetch schedule, fetch status, metadata, …).

Generator: Based on the data of crawldb, the generator selects best-scoring urls due for fetch and the segments directory is created.

Fetcher and CrawlDb Update: Next, the fetcher fetches the remote pages of the URLs on the fetch list and updates it to the segment directory. This step takes a lot of time.

Parser: The contents of each web page are parsed. If the crawl produces another extension to an already existing one, the updater adds the new data to the crawldb.

Inverting: The links need to be inverted before indexing. The fact that the number of incoming links is more valuable than the outgoing links is taken care off, similar to how Google PageRank works. The inverted links are saved in the linkdb.

***Indexing, Deduplicating and Merging**:* Using data from crawldb, linkdb and segments, the indexer creates an index and saves it. The Lucene library is used for Indexing.

4.3 Searching for test Queries

Now, we can search for tweets regarding the crawled database. The searching of the test query takes place in the following steps:

Stop Word Removal: From the given query the stop words have to be removed, because they do not contribute much to the searching procedure.

Query Segmentation: for a given set of words in a query, the search engine may not give proper results, so it searched for every combination of the words contained in the search query.

Merging: As discussed in the Query segmentation, the results for a given query may not be available; so the results obtained from the different combination of the words of the query text is merged to get one of the probable results of the search.

5 RESULT AND CONCLUSION

We submitted 37 results of the four query subtexts. The run submission was accepted in the category of semi automatic run. The results from the organizers after judging the submitted run is provided in Table 1.

Table 1. Evaluation Results of Semi-automatic Runs

Run Id	Precision @20	Recall @1000	MAP @1000	Overall MAP
nita_nitmz_1	0.0583	0.0046	0.0031	0.0031

The results are not encouraging, but considering the fact that we started from the scratch, we have much to learn. The different participating teams have employed different algorithms to extract the results. We would try to enhance our methodology for future research.

6 REFERENCES

[1] Vieweg S., L. Hughes A., Starbird K., Plaen L., 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*(Atlanta, GA, USA, April 10-15,2010). CHI '10. ACM, Atlanta, GA, 1079-1088. DOI=10.1145/1753326.1753486.

[2] Palen L., M. Anderson K., Mark G.,Martin J., Sicker D., Palmer M., Grunwald D. 2010. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference* (Swindon, UK, 13-16 April 2010). ACM-BCS '10, Swindon, UK.

[3] Verma S., Vieweg S., J. Corvey W., Plaen L., H. Martin J., Palmer M., Schram A., M. Anderson K., 2011. Natural Language Processing to the Rescue?: Extracting “Situational Awareness” Tweets During Mass Emergency. Association for the Advancement of Artificial Intelligence.

[4] Watanabe K., Ochi M., Okabe M., Onai R.2011. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*(Glasow,UK, 24-28 October 2011). CIKM '11, Glasow,UK, 2541-2544, DOI=10.1145/2063576.2064014.

[5] Lingad J., Karimi S., Yin J. 2013. Location extraction from disaster-related microblogs. *In Proceedings of the 22nd International Conference on World Wide Web(*Rio De Janerio, Brazil, 13-17 May*). WWW '13 Companion,* Rio De Janerio, Brazil, 1017-1020. DOI=10.1145/2487788.2488108

[6] Imran M., Elbassuoni S., Castillo C., Diaz F., Meier P. 2013. Practical extraction of disaster-relevant information from social media. *In Proceedings of the 22nd International Conference on World Wide Web(*Rio De Janerio, Brazil, 13-17 May*). WWW '13 Companion,* Rio De Janerio, Brazil, 1021-1024. DOI=10.1145/2487788.2488109

[7] Imran M., Castillo C., Lucas J., Meier P., Vieweg S. 2014. AIDR: artificial intelligence for disaster response. In *Proceedings of the 23nd International Conference on World Wide Web*(Seoul, South Korea,7-11 April). WWW '14 Companion, Seoul, South Korea, 159-162. DOI= 10.1145/2567948.2577034

[8] Anupam Jamatia,Amitava Das.Part-of-Speech Tagging System for Indian Social Media Text on Twitter. In *Proceedings Workshop on Language Technologies For Indian Social Media(SOCIAL-INDIA),* Pages 21- 28.

[9] Pinaki Bhaskar, Amitava Das, Partha Pakray, Sivaji Bandyopadhyay, Theme Based English and Bengali Ad-hoc Monolingual Information Retrieval in FIRE 2010, In *FIRE 2010*, Working Notes. [2010]

[10] Barman, U., Das, A., Wagner, J., and Foster, J. 2014 Code-Mixing: A Challenge for Language Identification in the Language of Social Media. In *The 1st Workshop on Computational Approaches to Code Switching*, EMNLP 2014, October, 2014, Doha, Qatar.

[11] Björn Gambäck, and Amitava Das.2016. Comparing the Level of CodeSwitching in Corpora. In the *10th edition of the Language Resources and Evaluation Conference (LREC),* 2328 May 2016, Portorož (Slovenia).

[12] Anupam Jamatia, Björn Gambäck, and Amitava Das. 2016. Collecting and Annotating Indian Social Media CodeMixed Corpora. In the *17th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING), April 3–9, Konya, Turkey.

[13] Kunal Chakma, and Amitava Das. CMIR:A Corpus for Evaluation of Code Mixed Information Retrieval of Hindi-English Tweets. 2016. In the *17th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING), April 3–9, Konya, Turkey.

[14] S. Ghosh and K. Ghosh. Overview of the FIRE 2016 Microblog track: Information Extraction from Microblogs Posted during Disasters. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016*, CEUR Workshop Proceedings. CEUR-WS.org, 2016.