International Journal of Latest Trends in Engineering and Technology
Vol.(12)Issue(1), pp.001-003
DOI: http://dx.doi.org/10.21172/1.121.01
e-ISSN:2278-621X

INFORMATION EXTRACTION FROM MICROBLOGS DURING DISASTER SITUATION: A REVIEW

Harshadkumar Prajapati[1], Dr. Vikram Kaushik[2]

Abstract: The objective of the paper is to review the literature, how the microblogs are useful in giving information in disaster situation. This paper throws lights on how microblogs are used in relief and rehabilitation work during disaster. The information extract by the systems are helpful to take appropriate decisions in the availability of various types of resources. This is important in the effective coordination of post disaster relief situation. The review has performed from the study of different research papers in the given context.
Keywords: Information Extraction, Microblog, Disaster, Gold Standard

1. INTRODUCTION

Information retrieval is the key area in the world of web now days, finding Material (usually text documents) of an unstructured nature which satisfies information need from within large collections [8]. User-generated content in Microblogging sites like Twitter is known to an important source of real time information on various events, including disaster events like floods, earthquakes and terrorist attacks etc. This review paper is the study of four different research papers which throw light on utilities of microblogs which are relevant and useful during post disaster situation using the various information retrieval techniques.

2. LITERATURE REVIEW

Moumita Basu, Anurag Roy, Kripabandhu Ghosh, Somprakash Bandyopadhyay, Saptarshi Ghosh [1] mentioned in their paper that microblogging site like Twitter and Weibibo are vital sources of information during disaster. They observed that important information is generally hidden from lots of conversional content. So, it is needed to develop automated IR methods to extract microblogs that contain specific type of situational information from number of microblogs posted. They collected a set of approximately 50,068 microblogs posted during Nepal Earthquake in April, 2015. The standard language model as implemented in Indri IR system [6] and word embedding based retrieval (word2vec [7]) model used to develop the dataset. With the consulting member of NGOs who works with the post disaster relief operations, they identified five critical topics on that basis information needs to be retrieved. As per the format used traditionally for TREC topics. The sample example is as shows in Table 1.

<num> Number:T1<title> What resources were available
<desc> Identify the messages which describe the availability of some resources.
<narr> A relevant message must mention the availability of some resource like food, drinking water, shelter, clothes, blankets, blood, human resources like volunteers, resources to build or support infrastructure, like tents, water filter, etc.

Table 1. Information needs as per the TREC format

The final gold standard contains the following number of tweets judged relevant to the five topics – T1 contains 589, T2 contains 301, T3 contains 334, T4 contains 112 and T5 contains 254.To retrieve microblogs relevant to the topics IR methodologies used for a given dataset. There were three stages in the retrieval (1) generating a query from the topic (2) retrieving and ranking microblogs with respect to the query (3) expanding the query and subsequently retrieving and ranking microblogs with respect to the expanded query. Language modeling and word embeddings methods used to extract from the topic text. It also found that query expansion can also be performed by Rocchio expansions methods and method based on word2vec. These two query expansion strategy identify different expansion terms for the initial query and ranking model. Based on the performance with manual and automatic query and two ranking models like language model of indri and word2vec-based model. The result shows that [1] word2vec based model gives better result in both the cases where ($p<0.05$).

Query Type	Ranking Model	Prec@20	Recall@1000	MAP@1000	MAP Overall
Manual	Word2vec	0.6700	0.6197	0.2343	0.2788
Automatic	Word2vec	0.4600	0.5591	0.1785	0.2242

[1] Research Scholar, Faculty of Computer Science, Sankalchand Patel University, Visnagar, Gujarat, India.
[2] Research Supervisor, Faculty of Computer Science, Sankalchand Patel University, Visnagar, Gujarat, India.

Table 2. Retrieval results using word2vec.

As per the experiments word2vec model captures the overall context of a tweet and it could differentiate between need and availability tweets, the same cannot be with language based approach (Indri Model).The retrieval performance for the expanded queries expanded by the Rocchio strategy or by word2vec strategy. With the initial queries (manual/automatic), two ranking model (indri/word2vec) and the two query expansion strategies (Rocchio and word2vec).The experiments results the retrieval performance is better for the initial queries than for the expanded queries because the manually generated queries were verbose and alredy contains the relevant terms results into saturation .The automatic queries and word2vec ranking retrieval better for the expanded queries than for the initial queries demonstrating the utility of query expansion in finding relevant terms missing in the initial automatic terms. For a given query and ranking model, Rocchio expansion strategy performs better than the word2vec expansion strategy.

Ribhav Soni, Sukomal Pal [2] focused on the creation of gold standard data for automatic retrieval of important tweets during disasters using various experiments on the gold standard data prepared in the FIRE (Forum for Information Retrieval Evaluation), they mainly found that gold standards data prepared in previous work missed many relevant tweets. They demonstrated a machine learning model which can help in retrieving the remaining relevant tweets by training an SVM model on a subset of the data and using it to get the most useful tweets in the entire data set.

This paper throws light on the ground truth annotations exercise missed upto four times as many tweets as were found. This represents a significant loss of information that could practically be very useful in a disaster situation. The accuracy of gold standard data is crucial for evaluation and comparison of retrieval system .It may lead to weaker system being ranked above better systems. As per the result they found that relevant tweets missing from the gold standard [5]. Based on subset of data SVM model used to retrieve tweets with highest confidence score. It found that averaged across all topics, only less than half of the relevant tweets among identified in the gold standard.

The seven information needs expressed as topics in TREC format used for retrieving relevant tweets. The gold standard involves three phases in which annotators independently search for relevant tweets, the tweets indexed using indri. All tweets identified by at least one of the annotators and relevance annotation finalize by mutual discussion and taken the top 30 results from each run and decide the relevance. They have taken a set of approximately 700 tweets randomly and judge their relevance for each of the seven topics. The result shows the number of relevant tweets identified annotation about 5 times of that identified in the gold standard. They trained machine learning models for automatic classification of tweets into topics for automatically retrieving most useful tweets that missed in the annotation. One tweet can be relevant to multiple topics so supervised machine learning models used for separation for each topic. Support Vector Machines (SVM) used for classification task, as it found the best classification models for text classification.

They concluded that the gold standard annotation exercise missed many relevant tweets even with three phase approach. They also highlighted some major reasons of happening so. First was of short and noisy tweets and second when the participating systems are large and diverse.

Du Xin, Wang Xiaoyu, Zhuang Ziyao, Qi Limin [3] developed a hybrid model for information retrieval from microblogs during disaster. In their paper they initially used classifier to differentiate the need-tweets and availability tweets from other tweets such as useless tweets and repeat forwarding tweets, then matching with need tweets with availability tweets performed. Both need and availability tweets available in different language, so this can be a task of information retrieval problem. Here various models like indri open-source retrieval tool and dirichlet language model used for retrieval purpose for various parameters set for SVM and LR. The average Map of LR algorithm is higher than that of two groups of using LibSVM.The precision of the three sets of values almost same but the first group of recall much lower than the other two sets. They deepen the study of machine learning and try to select more different features to filter and choose text content of informal occasions.

Saloni Baweja, Aniya Aggarwal, Vikram Goyal, Sameep Mehta [4] described the extraction of useful information like the need or availability of various resources and to find the tweets that express the need and availability of the same resources. Two tasks define in the experiments; first one is to identify the tweets indicating the need and availability of various resources like, electricity, medical aid, shelter, food, water, mobile, internet availability and second is to match the set of need tweets with appropriate availability tweets. The tweets classification process divided into three phases like preprocessing, feature selection and model selection. In preprocessing, the words start with the hashtags (#) and username mention with @ removed from given tweets. In feature selection, extract feature from the labeled tweets to train the binary classification model. Logistic regression algorithm used for the classification task. This model trained by the set of features extracted from non-duplicate need and availability tweets in the available tweets.

Authors concluded that a mix of linguistic and machine learning models to automatically identify the tweets which indicate the need and availability of resources in a disaster happened area. Tweet indicating the need of a particular resource and find the relevant tweets indicating its availability by calculating its cosine similarity, every tweet in this situation translated into a bag of noun words present in it.The relative sequence of words in the tweets play a significant role in improving the performance of the model which may be incorporated in features.

3. CONCLUSIONS

After reviewing the four research papers, we can conclude that microblogs are effective communication tool useful for handling disaster situations and can be used for relief and rehabilitation work. Various IR models and techniques have been used to extract the relevant tweets in the context, which have their own advantages and limitations. However, still there is a scope of improving value of precision, recall and MAP by developing better model.

4. REFERENCES:

[1] Moumita Basu, Anurag Roy, Kripabandhu Ghosh, Somprakash Bandyopadhyay, Saptarshi Ghosh, "Microblog Retrieval in a Disaster Situation: A New Test Collection for Evaluation" In ECIR 2017, 8th-13th April, 2017 in Aberdeen, Scotland UK.

[2] Ribhav Soni, Sukomal Pal, "Microblog Retrieval for Disaster Relief: How to Create Ground Truths?" In ECIR 2017, 8th -13th April, 2017 Aberdeen, Scotland UK.

[3] Du Xin, Wang Xiaoyu, Zhuang Ziyao, Qi Limin Rudra, "A Hybrid Model for Information Retrieval from Microblogs During Disaster", In working notes of FIRE 2017, Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10,2017

[4] Saloni Baweja, Aniya Aggarwal, Vikram Goyal, Sameep Mehta, "Automatic Retrieval of Actionable Information from disaster related Microblogs", overview of the FIRE 2017 track: Information Retrieval from microblogs during Disasters (IRMiDis)", Working notes of FIRE 2017- Forum for Information Retrieval Evaluation. CEUR Workshop Proceedings, Bangalore, India: CEUR-WS.org, Dec-2017.

[5] Ghosh, S., Ghosh, K. "Overview of the FIRE-2016 microblog track: information extraction from microblogs posted during disasters". In working notes of FIRE pp.7-10(2016).

[6] Strohman, T., Metzler, D., Turtle, H., Croft, and W.B, "Indri: A language model-based search engine for complex queries", In Proc. ICIA. Available at: http://www.lemurproject.org/indri/ (2004).

[7] Mikolov, T., Yih, W., Zweig, G, and "Linguistic Regularities in Continuous Space Word representation", In: NAACL HLT 2013(2013).

[8] Manning C., "Introduction to Information Retrieval", CUP.