Multilingual Microblog Summarization

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. Microblogging is prominent e-communication medium on which short story are updated by the user based on their personal matter and other happening or coming immediate information. The quantity of information is large and also most of the data are redundant or irrelevant because of their popularity. This paper provides effectual techniques for summarization of inside story on microblogs sites such as twitter. The twitter data is the incredibly huge amount of small story circulate by users related to occurring situation or events. This technique focuses on finding factual most similar information respect to the query and used the ranking function for retrieving top-ranked twitter data related to query. Apply similarity measure function on top-ranked Relevant Tweets for detecting novel Tweets and which minimize similarity and maximize dissimilarity of twitter data. And also utilize threshold based decision to find a summary of novel tweets.

Keywords: Real-time data, Social media, clustering, Multi-document summarization Information Search and Retrieval, Web-based services, Microblog

1 Introduction

Microblogging is popular E-communication medium on which user circulate their small story based on incident happening related to their personal or surrounding events. It's a simpler and faster than traditional forms of communication medium and become popular perpetually in every area.

Twitter is one of the most prominent microblogs sites at the present time. It allows users to posted short and persistent status not more than 140 characters are known as a tweet. Everyday people provided over hundreds of millions of tweets from different parts of the world. People can socialize and interact with each other on day to day basis.

The Twitter information inside the story depends on user attentiveness and change according to interest. Therefore, Twitter streams contain a large and diverse amount of information ranging from daily-life stories to the latest local and worldwide news and events [1].

In addition, the extensive amount of post has meant that it is nearly impossible to control and regulate the system. Twitter suffers from spam and irrelevant posts that

reduce its utility to some extent and most of it is unstructured containing duplicates and errors [2]. Millions of tweet updated so people have no time to visualize all those tweets. There is need to Provide the effective algorithm for search, extraction, and summarization of this information could create a coherent and comprehensive overview of the topic presented from several points of view [3]. So this paper finding real world most similar information respect to the query and used the ranking function for retrieving top-ranked twitter data relate to query. Apply similarity measure function on top-ranked relevant tweets for detecting novel tweets and which minimize similarity and maximize dissimilarity of twitter data. And also utilize threshold based decision to find a summary of novel tweets.

1.1 Challenges

- Limited content of a single post;
- Huge amount of posts (above 400 million updates circulate every day on twitter)
- Many posts don't give a significant, valid and useful information;
- User search information based on name entities such as organization, people, place, and events;
- Many of posts contain opinions and sentiments;
- Diverse people belonging to different region post tweet on the same event

1.2 Objectives

- Design and implement system to retrieve most relevance information From Twitter
- Do the Clustering of data and, to construct tweet summary of up to 100 novel tweets from the set of relevant tweet for a given interest profile

2 Problem Statement

Given set of tweets T and set of queries Q where T= {T1, T2, T3...Tn} and= {Q1, Q2, Q3..., Qn}

F is a function to summarization And Summary S= $\{s_1, s_{2,}...,s_n\}$ has formed from relevant tweet RT={rt1,rt2,…,rtn} here rti represent as relevant tweet for particular interest profile F: T -> S

A batch of top 100 ranked tweets per day per interest profile with any two tweets having a similarity of less than threshold sim(t1, t2) < Ts is used for the summary. dissim is dissimilarity of a set of tweets and sim is similarity of a set of tweets

$$\begin{matrix} \text{Max } \Sigma \text{dissim (T)} \\ \text{Min } \Sigma \text{sim (T)} \end{matrix} \quad (1)$$

3 System Architecture

In this Portion, we will identify a batch of top 100 ranked tweets per interest profiles. For high-level its results provide relevant and novel information for summarization purpose. Our system Architecture mainly contains four modules

3.1 Data Cleaning Module

We pre-process all raw tweets which performed lower casing and removing hashtags, hyperlinks, and punctuation. Also simply filtering these tweets which do not contain any keywords for each interest profile, and the remaining tweets are taking as candidate tweet collection for identifying possible relevant tweets of each profile.

3.2 Query Expansion Module

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.

3.3 Relevance Ranking Module:

We utilize the ranking function to measure the relevance between query and tweets. After that, all the tweets are ranked based on their relevance score and find the top ranked tweets related to interested profiles.

3.4 Novelty Detection Module:

When we obtain the top ranked tweet list after relevance ranking, we will have detect novelty for each tweet from, until we collect enough tweets to pushed into the summary. For novelty, we compared to tweets using Cosine similarity- function. This Module makes a threshold-based decision in which it considers a tweet with a similarity score above relevance threshold. A tweet is considered novel if its similarity score does not exceed a novelty threshold Tr compared to any of the pushed tweets, otherwise, the system ignores it. And pushed all tweets which similarity score less than the threshold into pushed tweet pool for making a summary.

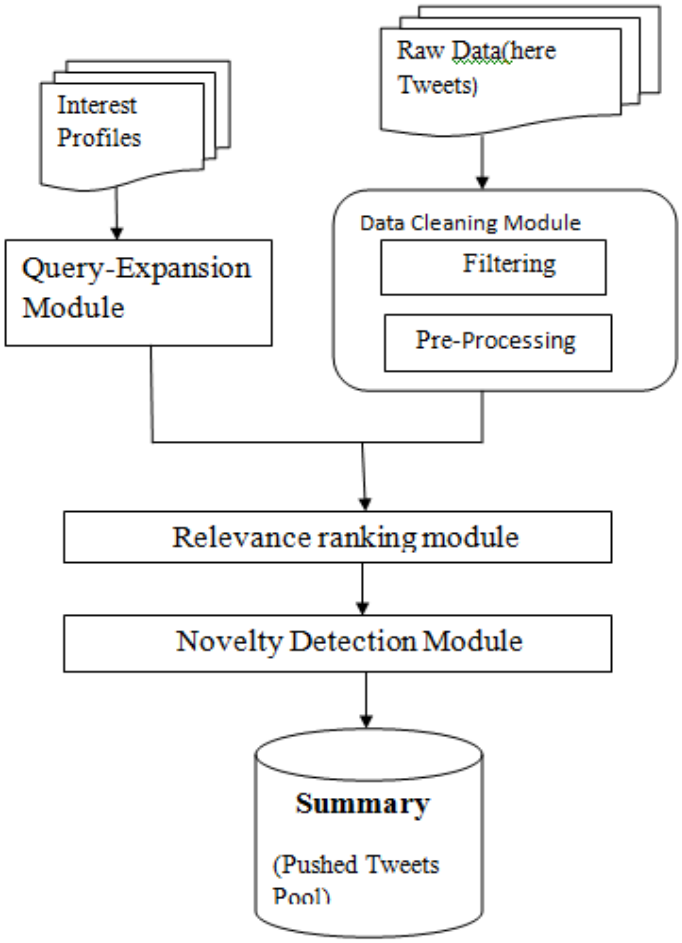

Fig. 1. Different System Components

4 Approach

In this portion, we represent as some strategy for summarization purpose. Based on this we used top-ranked relative data as an input. For minimize similarity and maximize dissimilarity of tweets we apply proposed algorithms to produce a summary of relevant tweets as output and in which also utilize decision-making function.

4.1 Cosine Similarity

Cosine similarity is a measure of similarity between two nonzero vectors of an inner product space that measures the cosine of the angle between them.

It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude.

The cosine of two none zero vectors can be derived by using the Euclidean dot product formula:

$$\text{Similarity} = \text{Cos}(\Theta) = A.B/\text{llAll.llBll} \quad (2)$$

.

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating orthogonality (decorrelation), and in-between values indicating intermediate similarity or dissimilarity

4.2 Jaccard Similarity

The Jaccard index, also known as the Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets and is defined as the size of the intersection divided
By the size of the union of the sample sets:

$$J(A, B) = |A \cap B| / |A \cup B|. \quad (3)$$
$$0 < J(A, B) < 1.$$

If A and B are both empty, $J(A, B) = 1$.

Jaccard distance measures *dis*similarity between sample sets:

$$J\delta(A, B) = |A \cup B| - |A \cap B| / |A \cup B| = 1 - J(A, B). \quad (4)$$

5 System Evaluation Result

Our system has been evaluated by the SMERP 2017 data challenge track. The evaluation score in terms of Recall (ROUGE-1), Recall (ROUGE-2), Recall

(ROUGE-L), and Recall (ROUGE-SU4) have been reported by the SMERP as .3471, .0622, .3233, and .1220 respectively and run type is Semi-Automatic.

Table 1. SMERP Level-1 Summarization Evaluation Result Table

SL No	Run-id	Run type	Recall (ROUGE-1)	Recall (ROUGE-2)	Recall (ROUGE-L)	Recall (ROUGE-SU4)
1	dcu_ADAPT_run3	Semi-automatic	0.5540	0.2436	0.5142	0.2864
2	USI_1	Semi-automatic	0.5187	0.2512	0.4796	0.2505
3	daiict_irlab_2	Semi-automatic	0.3515	0.1297	0.3254	0.1194
4	charusat_smerp17_1	Semi-automatic	0.3471	0.0622	0.3233	0.1220

6 Conclusion

In this paper present system architecture for real-time microblog summarizes techniques, Cosine Similarity and Jaccard Similarity. Apply relevance ranking model to rank candidate tweets and then we used strategies to measure novelty between tweets. And also I have makes a threshold-based decision for making summary which gives a better result. I will try to get the more accurate result using proposed algorithms and providing more training to the system.

.

References

1. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. Computational Intelligence, 31(1), 132-164 (2015)
2. Manning, C.D., Raghavan, P. and Schütze, H.: Introduction to information retrieval,Vol. 1, No. 1, p. 496.Cambridge university press, Cambridge (2008)
3. McDonald, R., April: A study of global inference algorithms in multi-document summarization. In : Proc. European Conference on Information Retrieval, pp. 557-564. Springer, Berlin Heidelberg (2007)