Microblog Processing : A Study

Sandip Modha
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat India
sjmodha@gmail.com.com

ABSTRACT

Sensing Microblog from retrieval and summarization become the challenging area for the Information retrieval community. Twitter is one of the most popular micro blogging platforms. In this paper, Twitter posts called tweets are studied from retrieval and extractive summarization perspectives. Given a set of topics or interest profiles or information requirement, a Microblog summarization system is desinged which process Twitter sample status stream and generate day-wise, topic-wise tweet summary. Since volume of the Twitter public status stream is very large, tweet filtering or relevant tweet retrieval is the primary task for the summarization system. To measure the relevance between tweets and interest profiles, Language model with Jelinek-mercer smoothing, Dirichlet smoothing and Okapi BM25 model are used. Behaviour of Language Model smoothing parameter λ for JM-smoothing and μ for dirichlet smoothing is also studied. Summarization is anticipated as clustering problem. TREC MB 2015 and TREC RTS 2016 dataset is used to perform experiment. TREC RTS official metrics $nDCG@10-1$ and $nDCG@10-0$ are used to evaluate outcome of experiment. A detailed post hoc analysis is also performed on experiment results.

KEYWORDS

Microblog, Summarization,Ranking,Language Model, JM smoothing, Dirichlet Smoothing

1 MOTIVATION AND CHALLENGES

Microblog become popular social media to disseminate or broadcast the real world event or opinion about the event of any nature. As on 2016, Twitter has 319 million active users across the world[1]. With this large user base, Twitter is the interesting data source for the real time information. On many occasion, it has been observed that Twitter was the first media to break the event. Many times, thousands of users across the world geography interact on same topic or interest profiles with diverse views. Following are the major challenges for Microblog summarization. Henceforth, topic or interest profile will be used interchangeability in rest of paper.

i) Since Twitter imposes limitation on length of tweet, it become very difficult for retrieval system to retrieve tweet without the proper context. So, tweet sparseness is the critical issues for the retrieval system.
ii) On Many topics, the volume of the tweet is very large. Most of the tweets are redundant and noisy.
iii) On Twitter, Some of the topics are being discussed for a longer period of time. They also diverted into many subtopics (e.g. demonetization in India, Refugee in Europe). It is very difficult to track topic drifting for an event. To track topic drifting, one has to update query vector by expanding or shrinking query term.
iv) Tweet often include abbreviation (e.g. Lol,India written as ind), smiley, special character, misspelling (tomorrow is written like 2moro). Tweet normalization is the biggest issue for microblog processing.
v) On many occasion, it has been found that native language tweets are in transliterated romanaized English

There are two cases for Microblog summarization [2] [7] (I)Online summarization or Push notification: novel tweet sent to user in real time where latency is important i.e. how fast we can deliver relevant and novel tweet to interested user. (II) Offline summarization (Email digest): At the end of day, system generates topic-wise novel and relevant tweet summary which essentially summarizes what happened that day. In offline summarization, latency is not important. In this paper,latter case is considered for the experiment

Summarization System should include relevant and novel tweet in summary. If there are no relevant tweet for a particular interest profile on a specific day, then this day is called silent day for that interest profile and summarization system should not include any tweet for that particular profile. If system correctly identify such silent day, then it should be awarded with highest score (i.e.1). If system include tweet in summary for an interest profile on silent day, it receive score 0

2 RELATED WORKS

Jimmy Lin and Diaz [2] [3] had introduced Microblog track since 2012 with objective to explore new IR methodology on short text. Mosad et.al.[1] has trained their Word2vec model using 4 years tweet corpus.They have used Okapi BM25 relevance model to calculate the relevance score. To refine the scores of the relevant tweets, tweets were re scored using the SVM rank package using the relevance score of the previous stage. Luchen et.al.[7] expanded title term each day with point-wise KL-divergence to extract 5 hashtags and 10 other terms. For relevance score, they have used unigram matching formula with different weight to original title terms and expanded terms. Our approach is similar to [6] but we have empirically tuned smoothing parameter for better results. In addition to this, we have also incorporated two level thresholds which are computed via grid search which control tweets to be part of the daily summary

3 DATA AND RESOURCES

TREC has started Microblog Track since 2012. In 2016 track was merged with temporal summarization and renamed as Real time summarization track [2]. An experiment is performed on TREC RTS 2016 dataset[5] and TREC 2015 dataset[3] to evaluate our system performance. Table 1 describe statistics of both datasets.

[1] https://en.wikipedia.org/wiki/Twitter

Table 1: TREC RTS 2016 and TREC MB 2015 Dataset description

Dataset Detail	TREC RTS 2016	TREC MB 2015
Total Number of Tweets	13 Mn	42 Mn
Interest Profiles for evaluation	56	51
Size of Qrels	67525	94066
Number of positive Qrel	3339	8233
Number of common Interest profiles between 2 Datasets	11	11
Tweet download duration	02-08-16 to 11-08-16	20-07-2015 to 29-07-15

4 PROBLEM STATEMENT

Given an interest profile IP =$\{IP_1, IP_2, ..IP_m\}$, and tweets T = $\{T_1, T_2, .., T_n\}$ from the Dataset, we need to compute the relevance score between tweets and interest profile in order to create profile wise offline summary S = $\{S_1....S_n\}$. Where S_i is the set of i^{th} profile-wise day-wise relevant and novel tweets. We can model profile specific summary as below.

S_i = $\{t_1, t_2, .., t_n\}$ where t_i,t_j ∈T

For given interest profile, Relevance score between tweet and interest profile is greater than specified silent day threshold T_s and relevance threshold T_r. In addition to this, these tweets should be novel i.e. similarity between all tweet of the summary should less that the novelty threshold T_n. if any tweet t_i is included in the summary for a particular profile on a given day then it should satisfy following constraints.

- Length of day-wise summary of Interest profile upto 100 tweets
- Sim(t_i ,t_j) ≤T_n ∀t_j ∈S_i (T_n = Novelty threshold)

5 PROPOSED METHODOLOGY

In this section, we describe our proposed approach to design a Microblog summarization system.

5.1 Query formulation from interest profile

Interest Profiles are consist of 3-4 word title, sentence long narrative and paragraph length narrative explaining detailed information need [2]. All the terms from title field and named entity from description and narrative fields are extracted to generate query. A dictionary is maintained to map named entity with abbreviated forms.

5.2 Tweet Pre-processing

Tweets and Interest profiles were pre-processed before calculating the relevance score.Non-English tweets are filtered using language attribute of tweet object.Non-ASCII characters are removed. Tweet having external URL embedded with text are expanded and text of external URL are merged with tweet text. Tweet without external URL and less than 5 tokens are filtered.

5.3 Relevance Score

To retrieve relevant tweets for a given interest profile, we have implemented language model with Jelinek Mercer, Dirichlet smoothing with parameters and μ respectively. In addition to this,we have also used BM25 ranking model to tank tweets. There are two types of days namely silent day and eventful day. An eventful day is one in which there are some relevant tweets for the given interest profile in a given day. In contrast, a silent day is one for which there is no relevant tweet for the given interest profile. The system should not include any tweet in the summary for that day for that particular interest profile. On a silent day, the system receives a score of one (highest score) if it does not include any tweet in the summary for that interest profile and zero otherwise. Detecting a silent day for a profile is a critical task for the summarization system. The Ranking function is defines as follow

$$F(IP, T) = P(IP|T, R = 1)$$

The above equation describe that if tweet is relevant how likely interest profile would be IP. The term P(IP|T) estimated by language model.

5.4 Summarization Method

To select the top relevant and novel tweets, we have designed a two level threshold mechanism. At the first level, for any interest profile on any day, if all the tweets ranked under this profile have scores less than silent threshold T_s, we consider this day as silent day and we will not consider any tweet in the interest profile's summary. We have empirically set silent day T_s using grid search. In the other case, where we get tweet scores greater than T_s, we normalize the tweet scores. We assign value 1 to tweet with highest score and assign relative values to the other tweets in the rage of 0 to 1. We include all tweets which values more than T_{r2} $\big($ normalized score of T_{r1} in the range of 0 to 1 $\big)$ and actual score T_{r1} in our candidate list and extract top k tweets. thee second level relevance threshold of T_{r1} and T_{r2} is also selected empirically using grid search.

5.4.1 *Novelty Detection using Tweet cluster.*

In this study, Microblog or Tweet summarization problem is anticipated as a tweet clustering problem. Once all the relevant tweets are retrieved, clusters are formed using jaccard similarity of tweet's text.Tweets having external URL or tweets having temporal feature in the text are given priority because such tweets are more informative than the tweet with only text and without external. we have used regular expression to extract temporal expression from tweet text.

6 RESULTS

To evaluate the performance of the system, Normal discounted Cumulative gain, nDCG@10 is computed for each day for each interest profile and is averaged across them [2].There are two variant namely: nDCG@10-1, nDCG@10-0. In nDCG@10-1 [8], on silent

Table 2: Result on TREC RTS 2016 with different ranking function using grid search

Ranking function	ndcg10-1	ndcg@10-0
Language Model with jm smoothing	0.3317	0.0998
Language Model with Dirichlet smoothing	0.3384	0.1116
Okapi BM25	0.3524	0.1131

day, system receive score 1 if it does not include any tweet in the summary for the particular interest profile and 0 otherwise. However, in nDCG@10-0, for a silent day, system receives gain zero irrespective of what is produced [2]. Our goal is to maximize the value of nDCG@10-0 and nDCG@10-1 jointly, which gives a wider picture, by tuning parameter and T_s in case of language model with JM smoothing and μ and T_s in case of Dirichlet smoothing.

While analyzing the evaluation metrics $nDCG@10-1$ and nDCG@ 10-0 on TREC RTS 2016 [2] [8], our system had failed in some of the interest profiles like RTS37(Sea World), MB265(cruise ship mishaps), MB365(cellphone tracking) where we could detect some of the silent days and had obtained some score in the nDCG@10-1 metric but did not score in the nDCG@10-0 metric. This is why we look at both the metrics while evaluating our system. The TREC RTS 2016 [5] organizers had considered nDCG@10-1 which adds gain on silent as well as eventful day as a primary metric to rank various teams. However, ndcg-0 which reflects how many relevant and novel tweets are part of the daily summary and does not add gain on silent day is also very important. In our analysis, it was observed that TREC RTS 2016 result[5] shows that empty run i.e. blank file with zero tweets scored $nDCG@10-1 = 0.2339$ which is more than average score of all the teams so is not a very accurate measure of judging. COMP2016 team [4] receive score $nDCG@10-1 = 0.2898$ and $nDCG@10-0 = 0.0684$. So it shows that 76 percent of the $nDCG@10-1$ score obtained by system is by remaining silent. In this experiment, we have tried to tune parameters which maximize nDCG-1 and nDCG-0 jointly. We believe that nDCG@10-0 is a very important metric which indicate that how much relevant and novel tweets were included in the summary. We report our best result with nDCG@10-1=0.3524 and nDCG@10-0=0.1131. without any sort of query expansion substantially outperform top team [4] in TREC RTS 2016[2]. Improvement in ndcg@10-0 shows that we have added more relevant tweet in interest profile summary which is better in a lot of senses

Table 2 shows system result with all standard ranking algorithm. Results show that all the ranking function perform in line with respect to each other, though Okapi BM25 model marginally outperforms language model. Our result on language model with Dirichlet smoothing and JM-smoothing outperforms result reported by [6]. The factor behind this outperformance is we have chosen parameter $\lambda = 0.1$ and $\mu = 1000$ and two level threshold mechanism. suwaileh et. al. [6] have set λ=0.7 and μ= 2000. Table 3 shows the 25 percent improvement in the results reported by top team of TREC RTS 2016 [4][5]. Table 4 shows system result on TREC MB 2015 Dataset [3]. Here thresholds are decided empirically not through grid search.

Table 3: Result Comparison with TREC RTS 2016 top team

metric	our result	COMP2016	QU	Blank run
ndcg@10-1	0.3524	0.2898	0.2621	0.2339
ndcg@10-0	0.1131	0.0684	0.030	0

Table 4: Result on TREC MB 2015

team	nDCG@10
our results LM with jm smoothing	0.2676
NUDTSNA	0.3670
CLIP CMU	0.2492

7 POST HOC ANALYSIS

In this section, we discuss comprehensive performance analysis of the summarization system from various perspectives. Since the massive dataset is used in the experiment, Tweet Selection or Tweet filtering is the primary task of the summarization system. Since Twitter restrict length of tweet,Tweet sparseness is the biggest challenge of the relevant tweet retrieval.

Interest Profiles are consist of 3-4 word title, sentence long narrative and paragraph length narrative explaining detailed information need [2]. The crucial part is how do we generate query from triplet as shown in Table 1. Luchen et al.[7] reported that title keyword play critical role in the retrieval. Our experiment also support these findings.

The objective of the summarization system is to identify all the clusters formed across the given period for all the interest profiles and should not include any tweet if the given day is silent for any interest profile. Performance of Summarization system depends upon 2 task (i) Relevant tweet retrieval (ii) Novelty detection across relevant tweet.

7.1 Interest Profile characteristics

During post hoc analysis, It has been observed that interest profile have different characteristic. Some of the interest profiles have spatial restriction. For example bus Service to NYC, gay marriage laws in Europe, job training for high school graduates US. For some interest profile such spatial restriction is not applied; user information is spread across the globe. E.g. emerging music styles, adult summer camp, hidden icons in movies and television

Generalized interest profiles have many silent day and interest profile with spatial named entity have more relevant tweet. Named entity play a very crucial role in relevant tweet retrieval. Some of the title of interest profile does not include NE so we extracted NE from narrative field and included in query. Interest profile or query which does not have named entity as query term perform very badly in result metric e.g. emerging music style.

7.2 Named Entity Linking Problem

Interest profile some time contain very generalize Named Entity. E.g. legalizing Medical Marijuana US and matching tweet contain a Named Entity Florida (Florida Medical Association to oppose medical marijuana ballot amendment in Florida). Due NE linking problem relevant tweet score less against the interest profile.

7.3 Named Entity Normalization

Due to limitation in length of tweet, Microblog user often writes named entity in abbreviated form. E.g. DEA(Drug Enforcement Agency). Though we have term like drug enforcement agency but we can not retrieve tweet with above normalize named entity.

7.4 Clustering Issues

Since Tweet summarization is multiple document summarization problem, each tweet along with external URL is considered as one document. Since Twitter is the crowdsourcing platform, many user report same event with different facts. So our novelty detection algorithm fails to cluster all following in tweet in same cluster.

T_1 : Woman Is Eaten Alive By A Tiger At A Safari Park
T_2 : Woman attacked by a tiger when she gets out of her car in a safari
T_3 : Horror at Beijing Safari World as tigers attack women who exited car, killing one, injuring another

7.5 Inclusion of Conditional event in Interest Profile

For the Interest profile like cancer and depression, our system performs very badly. Here user is looking for patient suffering from depression after diagnosed with cancer. It is very difficult to judge co-occurrence of both events in the tweet.

7.6 Inclusion of Sentiment in Interest profile

Interest profile, like Restaurant Week NYC includes sentiment and opinion or recommendation. Some of the tweet which are matching but does not include sentiment perspective are marked as non-relevant. In future we have to keep hidden feature like sentiment to increase the score of low score non-relevant tweet.

7.7 Hash-tag Identification

Hash-tag can be one of the features, for relevant tweet identification. Relevant Hashtag identification will increase the score of relevant tweet, e.g. key word is sea world and hash tag is #seaworld or self driving car the relevant hash-tag is #selfdrivingcar

7.8 Effect of Query Expansion

It has been observed that interest profile not having proper named entity, our system perform very badly in terms of evaluation metric nDCG-1 and nDCG-0 in majority cases. We also hypotheses that query expansion might work positively for these interest profiles. Our result shows that query expansion for such topic improvise the result nDCG-1 and nDCG-0. One can do query expansion bases upon interest profiles or case 2 case basis.

8 CONCLUSION

In this paper, we presented summarization system using language model with JM smoothing ,Dirichlet smoothing and Okapi BM25 model. Results show that All the ranking function perform in line with respect to each other. Though Okapi BM25 model marginally outperform language model. We have perform grid search to determine optimal silent threshold T_s and relevance threshold T_r.We have also identify smoothing parameter λ =0.1 for Language Model with JM smoothing and in the case of dirichlet smoothing $\mu = 1500$ for better results. We showed that by effectively choosing parameter λ and μ, we can outperform the result obtained by [6].

9 CURRENT WORK

TREC RTS metric give more emphasize to precision rather than recall. query expansion may include non relevant tweet in the summary thus it improve recall but precision decrease substantially and produce adverse effect on the results. Relevance thresholds are very critical for the summarization system for selection of tweet in the day-wise topic-wise summary. After doing careful analysis on TREC MB 2015 and TREC RTS 2016 dataset, we found that non relevant tweets have score more than relevant tweets in many occasions. It gives intuition for designing machine learning technique or deep neural network to estimate silent day threshold T_s S and relevance threshold T_r. As of now, we are working on following hypothesis.

H1: we can predict threshold for new dataset (TREC RTS 2016) from old data set TREC 2015 dataset.

Some of the interest profiles are common in both Datasets. Based upon this fact,we have designed following hypothesis.

H2: Irrespective of same topic or different topic, statistical features of the rank list can be exploited to predict silent day relevance threshold T_s and relevance threshold T_r

As of now, I am working on machine learning model for estimation of these thresholds for any Dataset downloaded from Twitter.

REFERENCES

[1] Mossaab Bagdouri and Douglas W Oard. 2015. CLIP at TREC 2015: Microblog and LiveQA.. In *TREC*.

[2] Luchen Tan Richard McCreadie Ellen Voorhees Jimmy Lin, Adam Roegiest and Fernando Diaz. [n. d.]. TREC RTS 2016 Guidelines. http://trecrts.github.io

[3] Yulu Wang Garrick Sherman and Ellen Voorhees Jimmy Lin, Miles Efron. [n. d.]. TREC 2015 Microblog Track: Real-Time Filtering Task Guidelines. https://github.com/lintool/twitter-tools/wiki/TREC-2015-Track-Guidelines

[4] Haihui Tan Dajun Luo Wenjie Li. [n. d.]. PolyU at TREC 2016 Real-Time Summarization. ([n. d.]).

[5] Jimmy Lin, Adam Roegiest, Luchen Tan, Richard McCreadie, Ellen Voorhees, and Fernando Diaz. 2016. Overview of the TREC 2016 real-time summarization track. In *Proceedings of the 25th Text REtrieval Conference, TREC*, Vol. 16.

[6] Reem Suwaileh, Maram Hasanain, and Tamer Elsayed. 2016. Light-weight, Conservative, yet Effective: Scalable Real-time Tweet Summarization.. In *TREC*.

[7] Luchen Tan, Adam Roegiest, Charles LA Clarke, and Jimmy Lin. 2016. Simple dynamic emission strategies for microblog filtering. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 1009–1012.

[8] Luchen Tan, Adam Roegiest, Jimmy Lin, and Charles LA Clarke. 2016. An exploration of evaluation metrics for mobile push notifications. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 741–744.