IITH at CLEF 2017: Finding Relevant Tweets for Cultural Events

Sreekanth Madisetty and Maunendra Sankar Desarkar

Department of CSE, IIT Hyderabad, Hyderabad, India
{cs15resch11006, maunendra}@iith.ac.in

Abstract. Retrieving relevant tweets corresponding to cultural events can be used in various applications like event reporting, event recommendation, etc. This type of retrieval is challenging due to short length of the tweet, noise, out of vocabulary words, abbreviations in the tweet. In this paper, we focus on the problem of retrieving relevant tweets related to given cultural event of a festival. We consider several factors like BM25, DFR, presence of artist name, relevant hashtag, festival name for finding the relevance of tweets to the event. We apply BM25 + DFR model to retrieve candidate set of tweets related to each event of a festival. We find the top hashtags for each event by exploring meta-attributes of an event. We re-rank the initial rank list from BM25 + DFR based on two strategies, namely, presence of the event meta-attributes (artist name, festival name, title, etc.) and the identified top hashtags in the tweet, and based on the timestamp of the event. We experimented on a subset of CLEF 2017 cultural microblog contextualization dataset. The experimental results show that the proposed method is able to put relevant tweets at the top of the retrieval list.

1 Introduction

There are three tasks in CLEF 2017 microblog cultural contextualization track, namely, content analysis, microblog search, and timeline illustration [2]. Festivals dataset is provided for all the tasks. There are 70,000,000 microblogs available in the dataset which is collected between May 2015 and October 2016. We focus on task 3, timeline illustration that aims to retrieve all relevant tweets related to each event of a festival provided in the topic queries [2]. Each topic is related to one cultural event. Here, one event means one occurrence of a show (theater, music, etc.). Same show can be performed on different days of the festival. Each such show is treated as a separate event. Each topic is described by meta-attributes such as id, festival name, title, artist (or band), start date, end date, and venue. We apply BM25, BM25 + DFR and several re-ranking methods for retrieving the tweets of cultural events. Re-ranking by event timestamp gives better performance as most of the tweets are posted at few days before, during or after the events scheduled time.

Rest of the paper is organized as follows: Section 2 describes the problem definition. Next in Section 3, details of the proposed method are presented.

Submitted runs for evaluation are described in Section 4. We conclude the work by providing directions for future research in Section 5.

2 Problem Definition

Here we briefly define the problem addressed in this paper: *Given an event E of a festival with its meta attributes title, artist name, festival name, start date, end date, and venue, retrieve all relevant tweets related to the event.* Such information is useful for attendees of festivals, for people that are interested in knowing what happens in a festival, and for organizers to get feedback [4]. Each event is represented as a topic. The following is the example of a topic in the festival.

```
<topics>
...
   <topic>
      <id> 1 </id>
      <title/>
      <artist> Anna calvi </artist>
      <festival> charrues </festival>
      <startdate> 16/07/15 - 18:45 </startdate>
      <enddate> 16/07/15 - 19:45 </enddate>
      <venue> Kerouac </venue>
   </topic>
...
</topics>
```

In the topics dataset provided, the *id* attribute ranges between 1 to 664, i.e., there are 664 topics in total for this task. The topic given in the above example specifies the live music show given by *Anna Calvi* in Vielles Charreus 2015 festival without any specific title, the *title* field is empty. The *artist* can be a single artist, a list of artist names, orchestra name, as they appear in the official programs of the festivals. The *festival* labels in the dataset are: *charrues* for Vielles Charrues 2015, *transmusicales* for Transmusicales 2015, *avignon* for Avignon 2016, *edinburgh* for Edinburgh 2016. *Startdate* is start date and time of the festival whereas *enddate* is end date and time of the festival. The *venue* is a string that corresponds to the name of the event location, given in the official program description.

3 Methodology

In this section, we describe our methodology to retrieve relevant tweets for given cultural events. There are three phases involved in our method, preprocessing the tweets, identifying relevant tweets, and re-ranking the tweets. Each of these phases is explained in detail in the following subsections.

3.1 Pre-processing the tweets

This is the first phase in our method. The dataset contains 70 million tweets [2]. We have observed that discussions about events often happen few days before or after the events scheduled time. So, we filter the dataset based on timestamps of the festival before the indexing step. For an event, while retrieving the relevant tweets, we consider only those tweets that were posted within ±15 days of the event. For example, if a certain event is happening on 15th June then we consider the tweets posted in the month of June. We assume that most of the tweets about this event are posted in this duration only. Although, there are tweets related to this event which are posted beyond this period, they are usually very less. This filtered dataset is used for faster indexing and retrieval of the tweets.

3.2 Identifying relevant tweets

This is the second phase in our method. We use two independent strategies to retrieve relevant tweets for a festival.

BM25: First, we apply BM25 (BM stands for Best Matching) also called as Okapi BM25 [7] scoring mechanism to retrieve the tweets related to an event. It is a probabilistic retrieval model. The ranking function used in BM25 is not a single function but a family of scoring functions, with a small change in parameters and components. Content matching to the given query is done by the BM25 algorithm. The ranking function of BM25 is as follows.

$$Score = \sum_{t \in q} log(\frac{N}{df_t}).\frac{(k_1+1)tf_{td}}{k_1((1-b)+b.(\frac{dl}{avg\_dl})+tf_{td}} \tag{1}$$

where N is total number of documents, df_t is document frequency of the term, tf_{td} is term frequency in document d, dl is document length, avg_dl is the average document length in the whole collection, k_1 is tuning parameter for controlling the scaling of term frequency, and b is tuning parameter for controlling the scaling of document length. We have used k_1=1.2 and b=0.75 in our experiments.

BM25+DFR: DFR (DFR stands for Divergence From Randomness) [1] is a method in Information Retrieval to retrieve documents related to a given query. There are three building blocks in this model: basic randomness model selection, first normalization, and term frequency normalization.

The first step in this model is to select basic randomness model to find the weight of a term in the document. Several possible randomness models are Poisson approximation of the binomial, Divergence approximation of the binomial, Bose-Einstein distribution, Geometric approximation of the Bose-Einstein, Inverse term frequency model, Inverse document frequency model, Inverse expected document frequency model [1]. Next step is to apply the first normalization. If a rare term does not occur in the document, then the probability of that term in the document is zero and is less informative. On the other hand, if a rare term occurred many times in the document then the probability of that term is high and is more informative. The risk component for the term is used in this step similar to [6]. This risk component is multiplied to the weight of the term described in the first step. Some of the first normalization techniques are ratio of two Bernoulli processes [1], Laplace's law of succession [3]. The last step is to normalize the term frequencies. Document length is considered for this normalization. The following equation describes the term frequency normalization.

$$tf\_normalized = tf * log(1 + c * \frac{avg\_dl}{dl}) \tag{2}$$

where $tf\_normalized$ is the normalized term frequency, tf is the term frequency, c is a parameter, avg_dl is the normalized document length, and dl is the document length. Equation 2 is referred as Normalization 2. BM25 can be calculated using DFR model. So, we use DFR version of BM25 in this method. We call it as BM25 + DFR. The components used in DFR model in our method are Inverse document frequency model for randomness, Laplace succession for the first normalization, and Normalisation 2 for term frequency normalization. The tweet set which is obtained from BM25 + DFR algorithm is denoted by *initial rank list*. The score obtained after applying BM25 + DFR model is denoted by $Score_{BD}$. Next, we will apply different re-ranking methods to re-rank the tweets from *initial rank list*.

3.3 Re-ranking

After pre-processing and identifying relevant tweets, the next phase is to re-rank the identified tweets as described in Section 3.2. The following are the different re-ranking mechanisms that we applied to re-rank the results of *initial rank list*.

- **Meta-Attributes**: The meta-attributes of an event are artist name, festival name, etc. We also come up with a list of top hashtags for each event. Top hashtags are obtained by finding the similarity between meta-attributes and hashtags. This re-ranking strategy checks for the presence of artist name in the tweet, festival name, and one of the top hashtags in the tweet. Frequency is also used along with similarity features. Tweets that are not having any of these features are ignored. For each tweet we compute a meta score as follows:

$$\begin{aligned} Score_m(tweet) = \delta(&\text{tweet contains artist name}) \\ &+ \delta(\text{tweet contains festival name}) \\ &+ \delta(\text{tweet contains any top hashtag}) \end{aligned} \tag{3}$$

  Here δ is a boolean function that returns 1 if and only if the predicate is true. We re-rank the tweets based on the combined score as:

$$Score_{BDM}(tweet) = Score_{BD}(tweet) + Score_m(tweet) \tag{4}$$

  where $Score_{BD}(tweet)$ is the score obtained by BM25 + DFR model.

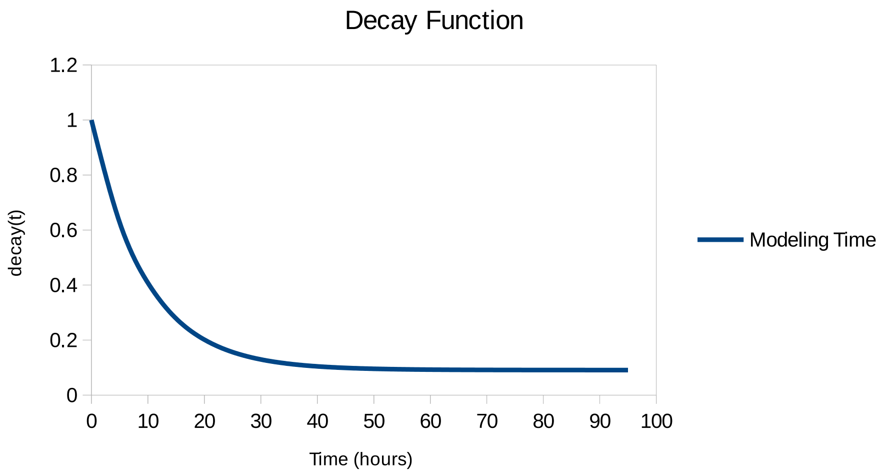

Fig. 1: Modeling Time

- **Time**: This method re-ranks the tweets based on the performance time of the event in the festival. We observe that some events are repeated on different days of the festival. In order to differentiate the tweets for repeated events scheduled on different days, the timestamp of the event will be helpful. Preference will be given to the tweets whose creation time is close to starting time of the show. Each tweet is assigned a time-based score to an event. If t is the absolute time difference between the event start time and the tweet creation time, then the $Score_{time}$ is computed as:

$$Score_{time}(tweet) = \frac{\gamma^t + \lambda}{1 + \lambda} \tag{5}$$

In our setting, we put $\gamma = 0.9$ and $\lambda = 0.1$. The value of $Score_{time}(tweet)$ is ranges between 0 and 1. The graph with different time values is shown in Figure 1. If the absolute time difference between event start time and tweet post time is less, then the time-based score of the tweet will be close to 1. Here, we have taken the time difference in hours. We used this function to suppress the tweets for which the tweet creation time is far away from event time and to give importance to the tweets for which the tweet creation time is close to the event time. We wanted this function to be non-linear in nature. If the time difference is small, then the importance dampens quickly, and after some time it will stabilize to some small but non-zero value. In this way, tweets that are posted far from the event time are still considered if the tweet content appears to be relevant to the event and get a chance to show up in the final ranking. We re-rank the tweets based on the combined score of BM25 + DFR, meta-attributes, and timestamp as:

$$Score_{final}(tweet) = Score_{BDM}(tweet) + Score_{time}(tweet) \tag{6}$$

4 Experiments

The dataset consists of 70 million microblogs [2]. Each microblog has the following attributes.

- id: unique id of the microblog
- from_user: author of the tweet (screen name)
- from_userid: unique id of the author
- iso_language_code: encoding of the tweet (en, es, fr, pt)
- source: interface used for posting the tweet (frequent tags: Twitter Web Client iPhone and Twitterfeed clients)
- profile_image_url: url of the profile image
- wday: week day
- created_at: tweet creation date
- time_s: quantitative variable (integer)
- time_ord: quantitative variable (integer)
- content: tweet content

The following is an example of a microblog.

```
    658495097328312321    loveethewayy    3318630764    en    IFTTT
http://pbs.twimg.com/profile_images/633606614843392000/ZdnClHU8_normal.jpg
Mon    2015-10-26    14840    1445832440    https://t.co/zf4cIQhXru
Le Festival dAvignon ou La passion thtre https://t.co/Aqjglu4QhP
via JulienGue
```

4.1 Runs Submitted

We have used Terrier IR platform [5] for indexing and retrieval of the tweets. We have submitted the following runs for task 3, timeline illustration, in CLEF 2017 microblog cultural contextualization track.

- Baseline1: Only BM25 is used in this method.
- Baseline2: BM25 + DFR combined run to retrieve the top ranked documents from each event of the festival.
- Meta Attributes: This run is performed using meta-attributes of the event like artist name, festival name. Top hashtags are also used as described in Section 3.3.
- MetaAttributes+Time: This run is performed based on Baseline2 + Meta-Attributes + timestamp of an event as described in Section 3.3.

The precision results of BM25, BM25 + DFR, Meta Attributes, and MetaAtrributes+Time are shown in Figure 2. The number of queries used for evaluation is 35. We observe that Meta Attributes precision values are higher than BM25 and BM25 + DFR models. Precision values of MetaAttributes+Time are greater than all other methods. This shows the importance of selected features, meta-attributes, and timestamp.

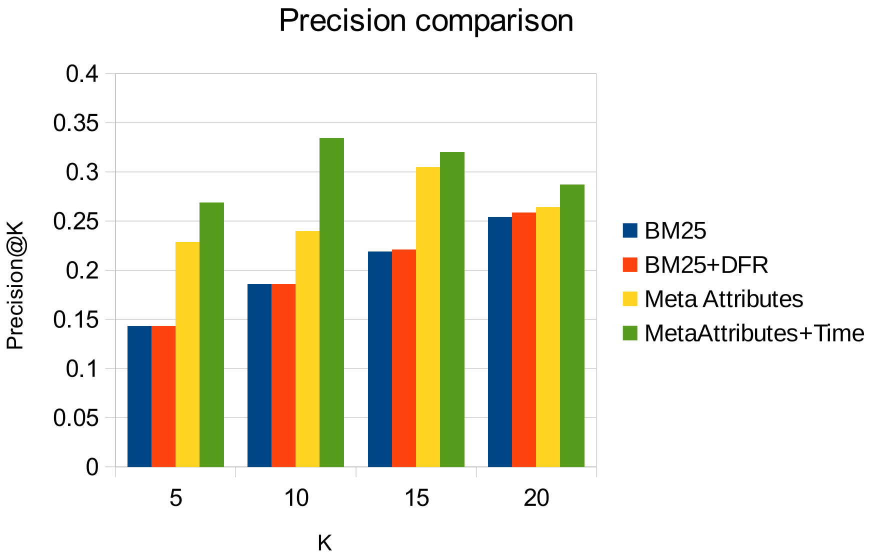

Fig. 2: Precision comparison of different retrieval methods

5 Conclusion

We assume that event to be a pre-planned event that is having context features like title, artist name, festival name, event start date, event end date, and venue. We have employed a method to retrieve relevant tweets for different events as part of time line illustration task, CLEF 2017 microblog cultural contextualization. We have used four different methods in this task. The first method use content matching for retrieving the tweets. The second method is DFR version of BM25. Remaining methods are based on re-ranking the *initial rank list*. We observe that re-ranking method which combines meta-attributes, and timestamp is performing better. For future work, we would like to see other methods which will further improve the recall.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS) 20(4), 357–389 (2002)
2. Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J.Y., SanJuan, E.: Clef 2017 microblog cultural contextualization lab overview. In: International Conference of the Cross-Language Evaluation Forum for European Languages Proceedings. Springer (2017)

3. Feller, W.: An introduction to probability theory and its applications, vol. 2. John Wiley & Sons (2008)
4. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 497–506. ACM (2009)
5. Macdonald, C., McCreadie, R., Santos, R.L., Ounis, I.: From puppy to maturity: Experiences in developing terrier. Proc. of OSIR at SIGIR pp. 60–63 (2012)
6. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 275–281. ACM (1998)
7. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., et al.: Okapi at trec-3. Nist Special Publication Sp 109, 109 (1995)