LIG at CLEF 2016 Cultural Microblog Contextualization: TimeLine Illustration based on Microblogs

Nayanika Dogra, Philippe Mulhem, Nawal Ould Amer, Lorraine Goeuriot

LIG at CLEF 2016 Cultural Microblog Contextualization: TimeLine Illustration based on Microblogs

Nayanika DOGRA [1], Philippe MULHEM [2], Nawal OULD AMER [1], Lorraine GOEURIOT [1]

[1] UGA LIG laboratory, MRIM group Grenoble, France
nayanika.dogra@e.ujfgrenoble.fr, nawal.ouldamer@imag.fr, lorraine.goeuriot@imag.fr
[2] CNRS LIG laboratory, MRIM group Grenoble, France
philippe.mulhem@imag.fr

Abstract. This paper presents the approach used by the LIG-MRIM research group to the participation of the task 3 (TimeLine illustration based on Microblogs) for the CLEF of Cultural Microblog Contextualization track. This task deals with the retrieval of tweets related to cultural events (music festivals) . For the content-based elements, we use the classical BM25 model [4]. Then, we diversify the results based on duplicate removal, using tf-based representations of tweets. In a third step, we apply optional re-ranking related to time-line, activity and popularity of authors of tweets.

CCS Concepts •Information systems → Information retrieval;

Keywords: tweet retrieval, diversification, reranking

1 Introduction

The goal of the Timeline illustration based on Microblogs subtask[1] is to provide, for each event of a cultural festival, the most interesting tweets. The Timeline Illustration Subtask focuses on two French music festivals (the "festival des vieilles charrues" and the "Transmusicales"), and the topics are all the live-events of one full day for each festival. Overall, there are 53 topics evaluated for this subtask. These topics are selected by the task organizers as live events corresponding to one day of each festival, and the goal is to retrieve relevant and diverse tweets related to each event. One example of topic depicts the show of *KhunNarin'sElectric* that took place at the *Transmusicales* the 03/12/15:

```
<topic>
  <id>1</id>
  <title>Khun Narin's Electric</title>
  <festival>Transmusicales</festival>
  <begindate>04/12/15-14:00</begindate>
```

[1] https://mc2.talne.eu/~cmc/spip/Tasks/task-3-timelineillustration-based-on-microblogs.html

<enddate>04/12/15-16:30</enddate>
</topic>

One of our goals four our participation to this retrieval task was to study the use of an information retrieval documents index as a basis for quasi-duplicates removal. Using such index allows to avoid complex partial string inclusions processes, and to use more simple overlap measures. Our overall approach is described in Figure 1, corresponding to the following organization of the paper. From the initial tweets set provided for the task, we filter (pre-process) the tweets the potentially relevant tweets as described in Section 2. Then Section 3 presents the content-based retrieval achieved. In a second step, a diversification process is achieved through a simple instance-based duplicate removal, as presented in Section 4. The reranking of the diversified tweets, in Section 5, is then performed using three different ways: timeline, tweet author activity, and tweet author popularity. We conclude this work in Section 7.

2 Pre-Processing of the official Tweet Corpus

The official corpus contains the tweets crawled during the months of July and december 2015.
Before indexing the tweet and processing the queries, we filtered the dataset to work on a subset of the official set of tweets provided. The filtering is based on timestamps, corresponding to the dates of the festivals, and text matching patterns (location or festival name for instance). The subset obtained consist on 243,643 tweets.

3 Content-based Matching

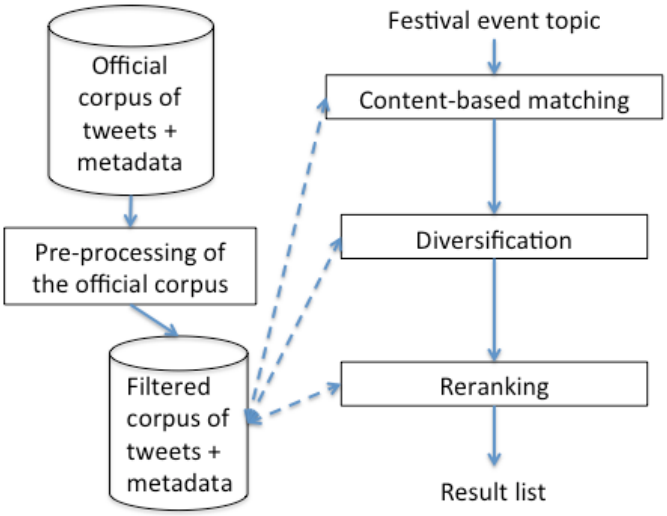

Figure 1: Overview of the query processing

The content-based retrieval is a simple process that uses the topic as the query, each query matched against the documents of the filtered corpus described in Section2. The content based retrieval uses BM25 [4] model.

4 Diversification

The second step of the query processing is dedicated to diversify the results. In the state of the art, several ways to diversify the results are proposed [1]. The authors of [2] mention that most of the diversification processes of the state of the art are achieved on after a first step of retrieval, and that is also our approach here. In the case of tweets, i.e., very short documents from which the content is very small, we chose to tackle this problem by removing duplicate tweets that correspond to *retweets*[2] In fact, our proposal does not limit the process to reteweets but to very similar tweets (that contain retweeted tweets). Here we propose:

- to keep the original tweet *t* when t and its retweets are in the result list;
- to keep the most relevant retweet of one tweet *t*, when several retweets are in the result list, but t is not retrieved.

Unlike we may think, this approach is not similar to achieving a flat clustering on the tweets, as we define a iterative process that goes from the top results to the last ones.
To avoid storing the original tweet in addition to their index, such filtering is not achieved on the initial text of the tweets, but directly their index that contain the (possibly stemmed) terms with their *t f* value). We use then an overlap function over the index of compared tweets, and a threshold above which the tweets are considered similar. If two tweets are considered similar, we keep one of these duplicates as described above. The result expected is then a short list of diverse tweets that describe the event.

5 Re-ranking

Having obtained content-based tweets, several ways of reranking them after the step 4 are explored:

1. No re-ranking (NO): The result of step 4 is directly given as an answer;
2. Timeline re-ranking (TIM): The result is re-ranked according to the creation date of the tweets. This kind of presentation allows the organizers of one event to pinpoint when something happened;

[2] One feature of Twitter is to allow users to “forward" (with or without alteration), or retweet, received tweets.

3. Social-based re-ranking: we defined two social based re-ranking functions as follows:
    - ACT: this re-ranking function is related to the activity of a tweet author. We assume that, the more active an author is, the more interesting are his tweets;
    - POP: this re-raking function is based on the popularity of tweet author. The underlying assumption being that the more the author is mentioned in tweets of the corpus, the more interesting his tweets are.

6 Experimental results

6.1 Parameters Settings

All our submitted runs are applied on the filtered corpus. The content-based retrieval uses the Terrier system [3], that implements BM25, using the default parameters (stoplist, Porter stemming, b = 0:75).
We tested three overlap values:

- the Jaccard overlap coefficient,
- the Szymkiewicz-Simpson cofficient,
- the Sorensen-Dice coefficient.

After some preliminary tests, for a and b coefficients the overlap threshold value is fixed to 0:75; and for the c coefficient, the overlap is fixed to 0.8. Because we do not have evaluation results for our runs, we only discuss the number of results obtained by these runs.

6.2 Runs submitted

We submitted the 7 following runs:

- RUN1: The content-only run, after the step 1 of the query processing described in Section 3. On average, each topic obtain a result list of 67 tweets;
- RUN2: Jaccard coefficient diversified-only run, obtained as the result of the step 2 of the query processing described in Section 4. On average, each topic obtained a 36 tweets long result list, so the diversity removes 45% results from the RUN1. Because the runs RUN5, RUN6 and RUN7 only reorder the results, they have the same result sizes;
- RUN3: Szymkiewicz-Simpson coefficient diversified-only run, obtained as the result of the step 2 of the query processing described in Section 4. On average, each topic obtained a 28 tweets long result list, so the diversity removes 59% results from the RUN1;

- RUN4: Sorensen-Dice Coefficient coefficient diversified-only run, obtained as the result of the step 2 of the query processing described in Section 4. On average, each topic obtained a 42 tweets long result list, so the diversity removes 38% results from the RUN1;

- RUN5: The results corresponding to the timeline reranking, *TIM*, as described in Section 5;

- RUN6: The results corresponding to the social activity-based reranking, *ACT*, as described in Section 5;

- RUN7: The results corresponding to the social popularity-based reranking, *POP*, as described in Section 5.

7 Conclusion

The participation to the subtask TimeLine illustration based on Microblogs of the Cultural Microblog Contextualization Workshop allowed us to define a comprehensive process for the retrieval of tweets. The pre-processing allows us to focus on a subset of the whole official set of tweets provided for the task. The content-based retrieval is a classical one. We used three variations of duplicate removal (diversification) methods that take into account the specificity of the tweets. We applied 3 ways to rerank the results in a third step of the query processing.
The impact of the pre-processing of the original corpus should be measured in the future, because it impacts the content-based matching, but also the activity and popularity values of tweet authors. Other variations of diversity algorithms also have to be studied, taking into account the specificity of tweets (especially their length, and their metadata), or even the choice of the kept tweet when we have duplicates.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5-14, New York, NY, USA, 2009. ACM.
2. C. Kuoman, S. Tollari, and M. Detyniecki. Using tree of concepts and hierarchical reordering for diversity in image retrieval. In *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*, pages 251-256, June 2013.
3. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform.

In *SIGIR'06 Workshop on Open Source Information Retrieval, (OSIR'06)*, 2006.

4. S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trecâĂŞ3. In *Overview of the Third Text Retrieval Conference (TREC-3), p*ages 109-126. Gaithersburg, MD: NIST, January 1995.