DMIR on Microblog Track 2011

Wen Li
Delft University of Technology
Delft, The Netherlands
wen.li@tudelft.nl

Carsten Eickhoff
Delft University of Technology
Delft, Neterlands
c.eickhoff@tudelft.nl

Arjen P. de Vries
CWI
Amsterdam, Netherlands
arjen@acm.org

ABSTRACT

In this paper we present our work on the Microblog Track of TREC 2011. We tried two methods to tackle the problem of tweet retrieval, namely EMAX and RTB. The first method EMAX is mainly based on the intuition that not only should retrieved tweets related to the keywords in given queries but also provide more information. This results in a ranking method based on self-information. Our second method RTB tries to incorporate the importance of recency along with relevance in microblog retrieval tasks. Therefore, we adapt portfolio theory to balance the relevance dimension and recency dimension. However, the evaluation results suggest no significant improvement from both two methods because of the short lengths of documents, the noisy and spam tweets and the re-ordering in recency. Meanwhile, we also present some ideas during the course of participation. By closely examining the judgments, we find that most of relevant documents are those containing a link to external resource and have a length of around 17 words, which is different from the statistics observed in the collection.

General Terms

Microblog Retrieval

Keywords

Twitter, self-information, portfolio

1. INTRODUCTION

The 2011 edition of TREC is the first to feature a dedicated track on microblog retrieval. As one of the participants, we submitted 2 runs on the JSON version of the provided data. In this paper, we will present our results as well as some further thoughts on the task and setting. The task of the track was to retrieve relevant tweets from a 14-day sample of Twitter. As a holder of a white-listed IP, Delft Multimedia Information Retrieval Lab uses the JSON version of the data to carry out all the experiments. In the following sections, we present our approaches EMAX and RTB alongside some general insights into the subject of microblog retrieval. To show the performance of our proposed system, we compared our runs with generic BM25 and TF-IDF implementations and a baseline provided by the organizers. Furthermore, we discuss some noteworthy observations, before concluding with an outlook on future directions in the domain of microblog retrieval.

2. MICROBLOG TRACK

2.1 Task

This is the first time for the TREC conference series to address the task of microblog retrieval with a dedicated track. The organizer released a set of 50 topics and all participants were asked to submit their retrieved documents for relevance judgment. For each query a proposed system should return up to 1000 documents for evaluation. A difference between this track from others is that queries are simulated as they are issued during sampling the collection. Therefore, no documents newer than the time of query should be retrieved. The top 30 retrieved documents of each query from each run were collected as a pool and judged by assessors arranged by the organizer. The judgment is based on a 3-level scale, namely high-relevant, relevant and irrelevant. During the judgment process, the assessors were allowed to use any resource to make a decision, e.g., by following links in tweets. However, all non-English and offensive documents were judged as irrelevant. The evaluation is done in a similar way. For each query in each run, all retrieved document are ordered by recency and then evaluated by traditional metrics, such as Precision @ 30 and R-precision.

In addition, participants were asked to submit at least one run without using any form of external (from outside of the collection provided) or future (from tweets with a timestamp newer than that of a query) information. Since microblog search is a realtime task, queries issued at different times request different parts of the document collection. This has to be reflected in the implementation of experiment systems, since the term statistics may vary across queries. One of the major contributions of our submission lies in providing the community with code to account for this special requirement.

2.2 Data

The document collection provided for all participants is a collection of tweets sampled during a two-week period from Jan 24 to Feb 8 of 2011, covering major events such as the Egyptian revolution and the US Superbowl. The document collection is crawled through Twitter's API following the track's ID list with our own multi-threaded crawler[1]. As a result, we obtained a collection of about 15 million tweets. It should be notice that all participants had to crawl the data themselves, as a result, it is likely that every participant uses a slightly different version of data. By matching the collection we crawled with the collection of relevant tweets

[1] Available at https://github.com/spacelis/tcrawl.

Table 1: Corpus Statistics

	Corpus	English
With Links	2,665,155	1,334,288
Users	5,228,689	2,554,641
Retweeted Tweets	1,672,912	762,907
Total	15,657,240	5,703,979

judged by official assessors, we found that we lost 21 relevant tweets, which is about 0.7% of the relevant. Since this task requires only English tweets, we apply an English language detection scheme to filter out all non-English documents. Concretely, we reject any tweet that contains more than 50% non-English terms. We used WordNet[2] to identify English words. After this step, 5.7 million English tweets remain in the collection. There is no relevant tweet lost in this step. A generic statistics for our collection before and after filtering is shown in Table 1. By comparing the statistics between English collection (indexed) and relevant collection, we have several observations discussed in Section 6.

3. OUR APPROACH

As a feature of microblogs, the limitation of text length forces users to write their messages concisely. Therefore, the tweets themselves usually are not able to carry as much information as normal documents in traditional retrieval tasks. As a matter of fact, the usual way to augment tweets with much richer information is to include hyperlinks pointing at external resources. With links, tweet writers can indirectly include any amount of information in any form (images, videos, etc.).

However, it also makes tweet retrieval more difficult since document relevance can hardly be decided by only looking at the tweet content. A solution to this problem can be to include the external evidence (resource) linked to in the tweet for ranking. Well-written web pages, images, video clips are common external evidence in Twitter. Among these types of evidence, web pages are more easy and promising to be integrated to a text retrieval system. However, a carefully composed web page could cost several hours to be put up on the Internet, which weakens the realtime feature of the tweets linking to it as they are delayed until the web page is available online. Furthermore, if we put too much weight on web pages linked to other than tweets themselves, retrieving tweets could degenerate to retrieving web pages which are manually collected and composed as tweets. This very problem of web retrieval is well studied in the past decades and has had a longterm dedicated track in TREC.

Meanwhile, other forms of external evidence is more interesting as they present information in multimedia dimension other than just text. For example, a user could tweet a picture showing the people marching and celebrating a festival or a short video shot at a football match when his favorite player gets a goal. This would be more interesting than just 140 characters. However, it is usually not an easy task to retrieve multimedia content. Therefore the only information we could easily use is the very descriptions of these images and video clips in tweets linking to them. At this stage of our research, we limit ourselves to the scenario of ranking tweets without using any external evidence.

3.1 EMAX

We assume a relevant and informative tweet to have two properties. It is i) related to the query and ii) carries sufficient information. For the first assumption, there are a number of state-of-the-art models, such as MB25 [3], TF-IDF [4, 2]. The second assumption is the key contribution of our approach to this ranking problem. Tweets themselves are usually short, particularly, they tend to be just one or two sentences. As a result, tweets usually carry a limit amount of information. Our approach tries to rank those carrying more information higher. Therefore, we use self-information to measure how much information a tweet (document) is loaded with. Self-information [1] is a concept from information theory which is used to measure how informative an event is in a probabilistic model. In our system, it expresses how much information a document d contains.

$$E(d) = -\sum_{t \in d} \log df(t)$$

where $E(d)$ is the information carried by document d, and $df(t_i)$ is the document frequency of term t_i.

Given a query, we first rank all documents by one of traditional retrieval models (namely, BM25 and TF-IDF) and then select the top 1000 documents as candidates. These top tweets are re-ranked in the descending order of their self-information. However, due to the fact that tweets commonly contain more informal terms than carefully edited resources, we expect higher entropy from them. In order to counter this effect, we filter out those words appearing less than 3 times in the collection.

3.2 RTB

Our second approach RTB (Relevance Time Balancing) is based on an application of the economical portfolio theory [5] and tries to balance multiple relevance criteria into one coherent ranking. As mentioned in the previous section, time is an important aspect in realtime retrieval tasks. Users are supposed in favor of recent relevant messages other than old well-known facts. Therefore, we propose a method to incorporating users' preference for recency by considering it a property of tweets. We combine the aspects of recency and topical relevance in the ranking by maximizing the target function $O(r)$ that takes into account the mean relevance $\mathrm{E}[R]$ and relevance variance $\mathrm{Var}(R)$ across multiple (in the current scenario 2) dimensions. The risk averseness parameter b can be used for further tuning towards particular user preferences. In the course of this work, it was statically set to 0.5.

$$O(r) = \mathrm{E}[R] - b\mathrm{Var}(R)$$

Similar to EMAX, we also apply RTB on top of several traditional retrieval models and also EMAX respectively.

4. IMPLEMENTATION

As we mentioned in Section 1, the temporal aspect is one of the important properties that makes microblog retrieval different from other ad hoc retrieval tasks. Traditionally, time-information is not directly involved in retrieval. As a result, all statistic figures (e.g., term frequencies, collection sizes, etc.) can be obtained statically at indexing time. However, for the realtime task required by this track, we know the concrete document collection to be included only at query time. During our work on this year's microblog

[2] http://wordnet.princeton.edu

Table 2: Evaluation

	Precision@30	R-precision
baseline	0.0986	0.1486
Emax	0.1007	0.1423
MB25	0.1041	0.1531
TFIDF32	0.2694	0.1828
RTB_EMAX	0.049	0.0472
RTB_TFIDF	0.049	0.0433
RTB_BM25	0.049	0.0433

track, we explored two ways to address this challenge. The first approach is straightforward. We index part of the collection according to the earliest overall query time and query the resulting index. For each subsequent query, we expand the index by the documents that were posted in the time between q_1 and q_2, and so on. Obviously, using this method, it takes very long to run 50 queries under different system settings. As a solution, we created a dynamic calculation of the term statistics used by the retrieval model. Our first version of implementation deals with document frequency which is typically used in many probability models. A patch to the Terrier search engine that accounts for the dynamic calculation of document frequencies can be downloaded from http://homepage.tudelft.nl/9y54n/terrier-3.5-realtime.patch.gz.

5. EVALUATION

In this section, we focus on the results of our unofficial runs, as unfortunately, at the time of official submission we had an implementation bug in our experimental system which led to a reverse ranking. After the submission deadline, we were able to correct for this. As suggested during discussion on the mailing list, we use both precision at 30 retrieved documents and R-precision as our evaluation methods.

To evaluate our first proposed method, EMAX, we compare it with a BM25 baseline and the official baseline. Figure 1 shows that there is no single method that can dominate the benchmark of P@30 and there is very little difference in the overall. Table 2 further emphasises this trend. The same can also be found for R-precision, as shown in Figure 2. Possible explanations for the insignificant differences are: i) all the methods rely on the same underlying components, term frequencies and inverse document frequencies, and ii) tweets contain too few words to properly characterize the topic they represent. A close look on the retrieval results suggests that EMAX tends to rank certain spam tweets higher since they have a lot of carefully selected words boosting their ranking. These tweets have three characteristics: i) they have many hashtags, ii) many repeated keywords and iii) they are usually very long. In the future, we would like to see whether we can apply spam filtering before ranking to resolve the problem.

Another interesting observation to be made is TF-IDF with a cut off at 30 scoring significantly higher than competing methods. Actually this can be considered a side-effect of the evaluation methodology, which we will address further in Section 6.

To evaluate our second method, RTB, we apply the balanced re-ranking algorithm on top of EMAX, BM25 and TF-IDF respectively. From the Figure 3 and 4, we can see that the results are worse than that before applying the balancing method. Because we did not consider a method to cut-off the retrieved list at a point higher than 1000 results, our approach does not lead to effective results under the evaluation. A possible reason is that not only recent relevant tweets get higher in the ranking but also recent irrelevant tweets are boosted into the top 1000 retrieved tweets. By the official evaluating method, those irrelevant but recent tweets may be then re-ordered at the top.

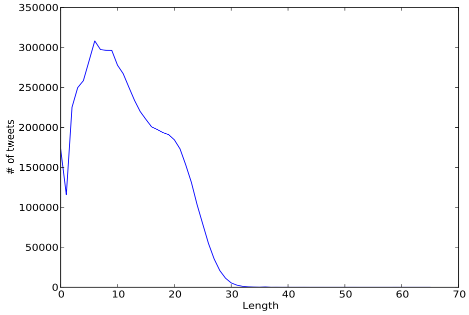

Figure 5: Length distribution in English collection

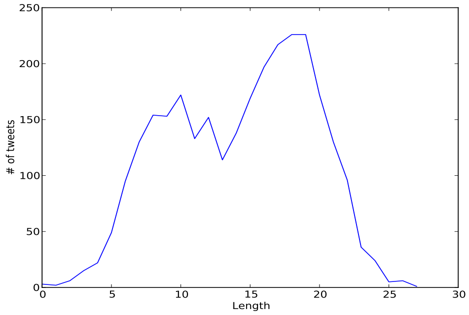

Figure 6: Length distribution in the relevant collection

6. DISCUSSION

As we mentioned previously, tweets are usually a sentence composed of several words because of the limitation of 140 characters. As can be seen from Figure 5, the typical length of a tweet is around 8 words. To our curiosity, we compare it with the lengths of relevant tweets judged by the assessors, which is shown in Figure 6. Besides the spike at around 9, relevant tweets are more likely to have a length of 17 words, which is about twice longer than normal ones. This may suggest that longer tweets are generally more probable to be relevant. It also supports our proposed assumption, as long tweets usually carry more information than short ones.

Inspired by the previous observation, we tried to find the difference in user dimension, which is shown in Figure 7 and Figure 8. However, there is no observable difference between the two collections in such user statistics. Meanwhile, we find an interesting user who contributes the highest number of relevant tweets. These tweets are judged relevant to query MB027 (reduce energy consumption), and the user is a broadcast account for a life hacking site on saving energy. This could be an interesting dimension to explore since it is reasonable that professional users usually provide more specific and informative tweets then others.

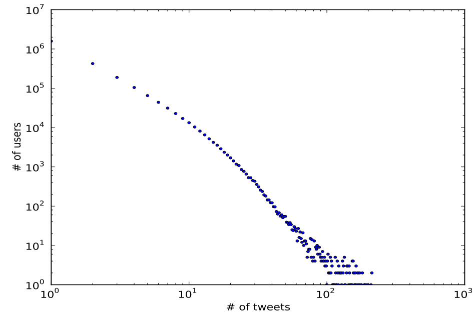

Figure 7: User distribution over tweet frequency in English collection

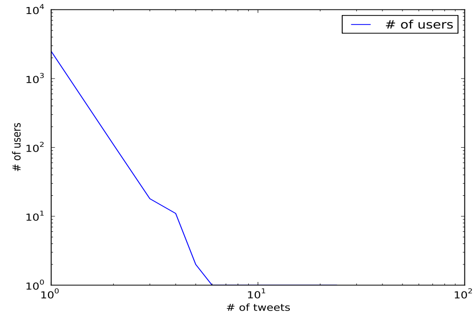

Figure 8: User distribution over tweets frequency in the relevant collection

By manual inspection of the evaluation results, we find that most of the provided queries are about events that would be expected to show up in news reports. We also checked the judgments used for relevance evaluation, and find that most of relevant tweets (94% of highly relevant tweets and 81% of all relevant tweets) contain links to external resources. Meanwhile, in the collection of judged irrelevant tweets, the proportion is only 53%, as shown in Table 3. It seems that tweets with links are more likely to be relevant in general.

This observation may be biased to the strategies employed by the majority of participants. It could be the case that some systems in favor of external evidence happen to be the ones who retrieve most relevant tweets and, consequently, influence the resource selection. It could also be influenced by assessors' perception of the relevance criteria and many queries are regarding news events. However, we could not provide hard evidence to support our suspicions. In spite of those, external evidence is assumed to greatly support the task, as they usually provide more relevant information to given queries.

The evaluation scheme use in the Microblog Track is a bit different from other tracks since recency becomes an important dimension to evaluate. The official evaluation with ordering in recency before traditional metrics is a natural way to reflect the demands on latest information while being compatible with default TREC evaluation scheme. Considering the set-based nature of the evaluation, a good system would actually separate the relevant from the non-relevant tweets, and only return the assumed relevant set. Our strategy to return 1000 tweets for every topic does not satisfy that evaluation design, and to test this we constructed the TFIDF30 run. Here, we only take the top 30 tweets ranked by TF-IDF and feed them to the evaluation system and produce better results than the baseline as shown in Table 2. However, identifying a good cutting point is usually hard in IR domain, and we would like to investigate this problem further in the future.

Table 3: Links in Relevant tweets

	With links	Total	Proportion
High relevant	527	558	94%
All relevant	2330	2864	81%
Not relevant	20200	37900	53%
Judged	4350	40764	11%

7. CONCLUSION

In this paper, we present our proposed methods for microblog retrieval and discuss some ideas that influenced our TREC participation. Generally, our proposed methods were not able to significantly outperform the baseline. The reasons probably are i) there are too few words in tweets which makes characterizing their topic hard for term-based models. ii) Spam tweets can get higher ranks in EMAX as their content is designed to cheat. iii) RTB tries to balance recency and topicality in retrieved documents which does not fit the design of the official evaluation methods. As a result, we only achieve performance similar to the baseline. The inspection of queries and the official judgments shows that tweets have a length of 17 words and with external links are more likely to end up in the relevant set. In the future we may experiment with length and URL presence priors.

8. REFERENCES

[1] `http://en.wikipedia.org/wiki/Self-information`.

[2] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

[3] S. Robertson, S. Walker, and M. Beaulieu. Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive. In *the Seventh Text REtrieval Conference*, 1998.

[4] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[5] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proc. of the Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR)*, 2009.

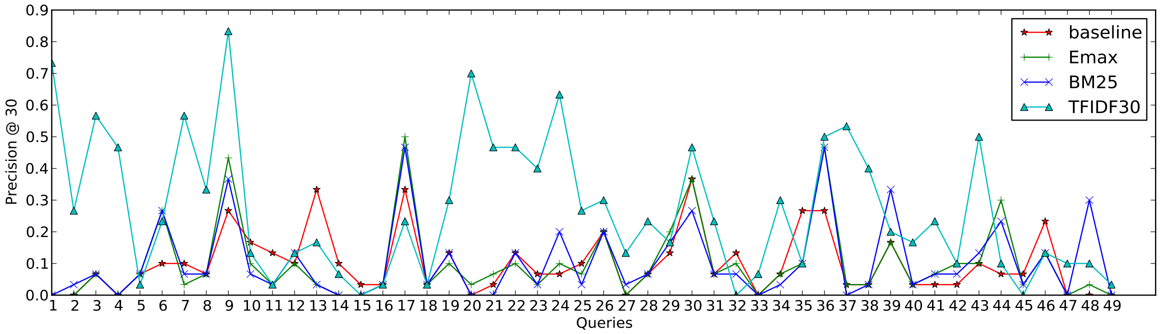

Figure 1: Precision@30: Comparison between Baseline, EMAX, BM25, TFIDF30

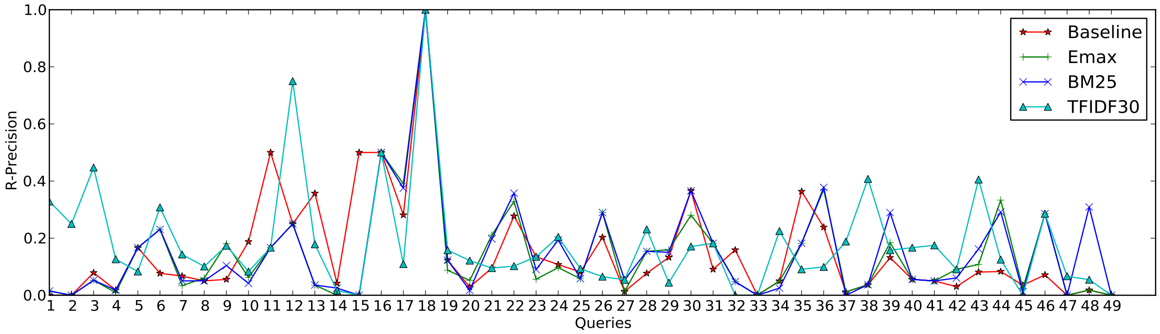

Figure 2: R-Precision: Comparison between Baseline, EMAX, BM25, TFIDF30

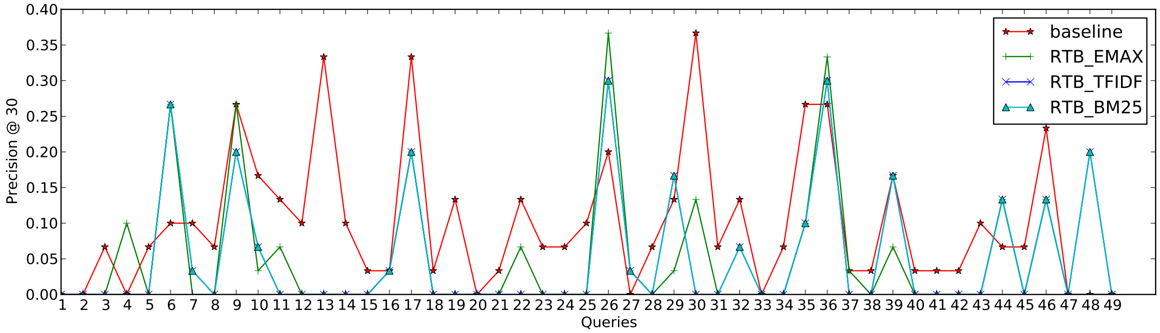

Figure 3: Precision@30: Comparison between Baseline, RTB_EMAX, RTB_BM25, RTB_TFIDF

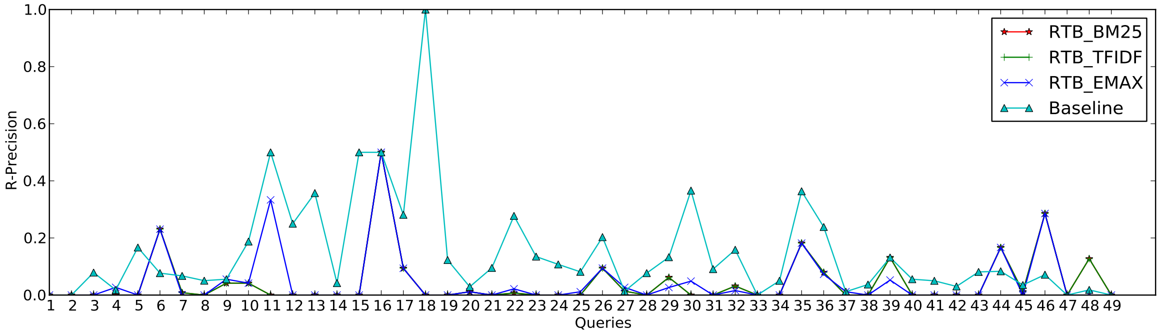

Figure 4: R-Precision: Comparison between Baseline, RTB_EMAX, RTB_BM25, RTB_TFIDF