Tie-breaker: A New Perspective of Ranking and Evaluation for Microblog Retrieval

Yue Wang, Jerry Darko, and Hui Fang

Department of Electrical and Computer Engineering
University of Delaware
140 Evans Hall, Newark, Delaware, 19716, USA
{yuewang,jdarko,hfang}@udel.edu

Abstract. Microblog retrieval is the key tool that enables users to access the relevant information from the enormous tweets posted on social media. Due to the differences of the tweets and traditional documents, existing IR models might not be the optimal choice for this problem. In this paper, we aim to introduce a new idea, i.e., tie-breaking, and discuss its implication in ranking methods and evaluation measures for microblog retrieval.

1 Introduction

Microblog has become an important information source in daily life. However, since the tweets are different from the traditional document in many aspects, the existing retrieval models might not provide the best performance in this domain specific problem. One feature of microblog is the length is shorter than the documents. Therefore, it is common that a retrieval function would assign the same score to multiple documents. When this happens, it creates ties in the retrieval results. These ties are interesting from both evaluation and retrieval perspective.

First, existing IR evaluation measures often assume that each document has a unique relevance score. As a result, the ties among documents with the same score are often randomly broken based on arbitrary factors such as document IDs. However, a more reasonable approach would consider all possible rankings for the tied documents, and then consider the performance range as well as the expected performance. As an example, when using IDF to rank tweets in Microblog12 data, the performance range would be 0.289 to 0.497 with the average performance of 0.363 in terms of MAP. When using Okapi BM25 to rank, the performance range is 0.322 to 0.333 with the average performance of 0.324. It is clear that IDF might be a better choice than Okapi BM25 since it has a greater potential to perform better because of the larger upper bound. However, existing evaluation methodology cannot detect such potential. We think that this observation and possible solutions to this problem are worth discussions.

Second, existing IR models often combine multiple retrieval signals in a very complicated way such as in pivoted or Okapi BM25 method. The complex equation mainly comes from the importance of balancing different retrieval signals. However, in the context of Microblog retrieval, it seems that many retrieval signals (e.g., TF and length normalization) become less important due to the special characteristic of tweets. Thus, an interesting research question would be whether it is still necessary to combine retrieval signals in such complicated ways, and whether there is any simple yet effective method.

In this years track, we applied a recently proposed tie-breaking strategy [1] for ranking tweets. Our experiments on the 2011 and 2012 collections have shown that this method is more effective than the traditional retrieval models such as Okapi BM25 and pivoted. For example, the performance (MAP@1000) of the tie-breaking method on Microblog 2011 is 0.374 (compared with 0.365 for pivoted and 0.323 for Okapi), and the performance on Microblog 2012 is 0.230 compared with 0.215 for pivoted and 0.192 for Okapi. In addition to the existing framework of tie-breaking, similar to the work proposed by Wang et al. [2], we also extend it with two commonly used strategies, i.e., query expansion and document expansion. The results show that the performance of the expansion on tie-breaking could improve the performance.

Table 1. Comparing the strength of the best implementation for signals

Retrieval signal	Implementation	$\mathbf{MAP_b}$	$\mathbf{MAP_w}$	$\mathbf{MAP_a}$
TF	$\frac{c}{c+1}$	0.3906	0.2306	0.2842
IDF	$log(1+\frac{N}{df})$	0.4969	0.2886	0.3631
DL	$\frac{1}{dl}$	0.0319	0.0220	0.0234
NOF	$log(NOF)$	0.0747	0.0562	0.0642

2 Tie-breaking on Microblog retrieval

2.1 Framework of tie-breaking method

Tie-breaking is a novel retrieval method, which tries to combine different retrieval signals in a simple but effective way[1]. The basic idea of tie breaking is to first prioritize all the signals and then apply one at a time to rank the documents. The signals are iteratively applied to break the ties generated by the previous signals, until all the signals are applied.

One of the key points of tie-breaking is the retrieval signal selection. Following the previous work[1], we also applied the term frequency (TF), inverse document frequency (IDF), and document length (DL) as the basic retrieval signals. In addition to that, number of follower (NOF) is also used in our experiment because it represents the endorsement from the online community. The number of follower a user has indicates how reliable and informative the user could be. We do not use other singles such as number of retweets since they are not available in the data set we get from API.

The next step is to decide the implementation of the retrieval signals. We tried different implementations for each signal on microblog 2011 collection data set, and report the best one in Table 1. We measure the effectiveness of the signals in the same way as described in previous work [1]: The strong signal should have (1) greater potential (i.e., larger MAP_b) and (2) better expected performance (i.e., larger MAP_a). Each implementation is reported with the best, worst, and average performance of MAP (e.g., MAP_b, MAP_w, MAP_a).

The last step is the identify the order of these signals. Intuitively, the strongest signal should be applied first in order to leave a larger space for following signals to improve the performance. As it is clear from Table 1, IDF has the best performance, while the TF is worse then IDF but better than the DL and NOF. DL is the weakest signal among all of these four. Therefore, we applied the signal in the following order: $IDF \bigoplus TF \bigoplus NOF \bigoplus DL$.

We compare the performances of the Tie-breaking method with two state-of-art methods: (1) **Okapi**: Okapi BM25 method; and (2) **Pivoted**: Pivoted document length normalization method. The results are shown in Table 2. All methods are trained on microblog 2011 collection, and tested on 2012 collection. It is clear that the tie-breaking method is more effective than the traditional retrieval methods.

Table 2. Comparing the performance of Tie-breaking and baseline methods.

	2011(Training)		**2012(Testing)**	
Method	**P@30**	**MAP**	**P@30**	**MAP**
Okapi	0.3782	0.3231	0.3029	0.1918
Pivoted	0.3964	0.3652	0.3349	0.2148
Tie-breaking	0.4204	0.3743	0.3684	0.2299

3 Extension of tie-breaking method

The tweets are shorter than the traditional documents. It is more important to bridge the vocabulary gap between the query and the tweets in microbolg domain. Following the previous works [2], we discuss how to extent the tie-breaking method to solve this problem.

3.1 Document expansion

Other than the query expansion, we could expand the document to bridge the vocabulary gap of the tweets and the query. As the length of the tweets are limited, it is common that users would insert a

URL, which links to the webpage that contains more information about the content of the tweets, in order to enrich their expressions. Thus, it is safe to take the content of the URL as the expansion of the original tweets.

To utilize the information from the expanded documents, we could extract retrieval signals from the expanded index, and then take them as the supplement signals to the signals from the originals tweets and perform the tie-breaking on all of these signals. Specifically, we utilized the TF, IDF, and DL signals from the expanded index. Together with the signals from the original tweets, we applied tie-breaking of all of these signals in the order $IDF_{orig} \bigoplus TF_{exp} \bigoplus TF_{orig} \bigoplus IDF_{exp} \bigoplus NOF \bigoplus DL_{orig} \bigoplus DL_{exp}$ based on preliminary results.

3.2 Query expansion

Query expansion is a common way to bridge the vocabulary gap by introducing more query terms to the original query. One typical method is pseudo relevance feedback, which extract terms from the initial retrieval results to expand the original query. Inspired by this, we propose the following method to extent the tie-breaking method with query expansion. We first perform the tie-breaking method with all the signals to get the initial results. We then select the top N terms from the top k documents as the expansion terms. Combine these expanded terms with a weight α to the original query. At last, we could use the tie-breaking method on the combined query to retrieve again.

4 Experiment

4.1 Experiment setup

We leveraged the API[1] provided by the originator to crawl the data set. Specifically, we crawled two sets of data. In one set we only use the original query to search the API and set the number of hit to 10,000. The other data set is crawled with both the original query and the query suggestions from Yahoo!, and the number of hit for each query is also set to 10,000.

We applied two methods to filter the non-English before we build the index on these tweets. We did not use the language field in the returned tweets as all of the tweets have the "null" as the language tag. Instead, on the tweets that are crawled only with the original query, we look at the ASCII code of the content of the tweets. If the ASCII code for every character of a word is in the range [32,127], we identify this word is a English word, and we keep the tweets that at least 50% of its contents is English words. Other tweets are removed from the collection. For the tweets that are crawled with the query suggestion from Yahoo!, we applied the *lang_id.py* package [3] to identify the language of each tweets. Only the tweets that are judged as English are kept.

In addition, the retweets are removed from the collections, while the non-retweets part from the partial retweets are kept. Due to the nature that the tweets are sensitive to the time, we built the index for each query in order to ensure that no tweets after the query time in involved. We applied the Krovetz stemming to the tweets, and no stop words is removed. We also crawled the links in tweets as the expansion for the original tweets. We trained our method on the 2011 collection in terms of P@30, and then tested it on 2012 collection. The best methods are submitted for the 2013 Track.

4.2 Experiment results

We submitted four runs to participate this year's track. The results are reported in Table 3:

(1) **UDInfoMINT:** We applied the query expansion method with tie-breaking as the retrieval function for this run. We use the index that crawled with the original tweets in order to ensure that no external and future information is used. All the following runs utilized the index that are crawled with both original query and query suggestion.

[1] https://github.com/lintool/twitter-tools/wiki/TREC-2013-API-Specifications

Table 3. Comparing the performance of different runs

	2011(Training)		2012(Testing)		2013	
Run ID	**P@30**	**MAP**	**P@30**	**MAP**	**Rprec**	**P@30**
UDInfoMINT	0.3816	0.2994	0.3048	0.1894	0.2654	0.3978
UDInfoMB	0.4701	0.3964	0.4207	0.2613	0.3285	0.5022
UDInfoMTB1	0.4361	0.4028	0.4201	0.2532	0.2873	0.4778
UDInfoMTB2	0.4449	0.3707	0.3655	0.2210	0.2614	0.4061

(2) **UDInfoMB:** We use only IDF as the retrieval function and perform pseudo relevant feedback. In addition, we also applied the IDF as term weighting on the expansion index. The two results are combine based on a normalized score as described in [4] with the weight on the original index set to 0.8.

(3) **UDInfoMTB1:** We applied the tie-breaking with all signals method as described in Section 3.1.

(4) **UDInfoMTB2:** We applied the tie-breaking with query expansion method as described in Section 3.2, i.e., tie-breaking plus the pseudo relevant feedback method.

4.3 Discussion

From the results, it is clear that the tie breaking method could out perform the traditional retrieval even apply the query expansion method (i.e., UDInfoMINT). On the training set, extensions of tie-breaking outperform the basic framework of tie-breaking, and the performances are comparable with the traditional retrieval method with query expansion and document expansion. However, the results of the proposed methods on this year's track are not as good as they are on the training sets. Comparing the query expansion and document expansion for the tie-breaking, the query expansion is even worse. We think the reasons of the poor performance could be as follow. First, the order of document expansion on tie-breaking may be sensitive to the expansion collection. The current order is trained based on the 2011 collection. When we crawl the links in the 2011 collection, a large number of them are not available due to multiple reasons. Therefore, the crawled expansion documents may different from the document collection crawled for 2013 tweets in terms of importance of the retrieval signals. Second, the query expansion for tie-breaking is worse than other method probably caused by the limitation of tie-breaking method, which assumes that every query term is important and may not perform well for long queries. Last, we want to point out the UDInfoMB is a strong baseline to beat as it involve both the query expansion and document expansion at the same time, while the tie breaking method only utilize one of these two.

5 Conclusion

Tie-breaking is a novel retrieval method for combining the retrieval signals in a natural but effective way. In this year's track, we explored the extended framework of tie-breaking and its performance on microblog retrieval. The experiment results show that The basic tie-breaking framework is more effective than the traditional retrieval method in tweets retrieval. Both query expansion and document expansion of tie-breaking has the potential to improve the performance, while document expansion seems more reliable than query expansion for tie-breaking. This shows the limitation of the current expansion methods. In the future, we plan to study how to better incorporate the expanded signals to the tie-breaking framework. In addition, we will explore more signals, such as the temporal-related signals, to the tie-breaking framework.

References

1. Wu, H., Fang, H.: Tie breaker: A novel way of combining retrieval signals. In: Proc. of ICTIR'13. (2013)
2. Wang, Y., Wu, H., Fang, H.: An exploration of tie-breaking for microblog retrieval. In: Proc. of ECIR'14. (2014)
3. Lui, M., Baldwin, T.: langid.py: an off-the-shelf language identification tool. In: ACL. (2012) 25–30
4. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM TOIS (2011) 8:1–8:34