Learning to Rank Microblog Posts for Real-time Ad-hoc Search *

Jing Li[1,2], Zhongyu Wei[3], Hao Wei[1], Kangfei Zhao[1]
Junwen Chen[4] and Kam-Fai Wong[1,2]

[1] The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2] MoE Key Laboratory of High Confidence Software Technologies, China
[3] The University of Texas at Dallas, Richardson, Texas, USA
[4] Tencent, Nanshan District, Shenzhen
{lijing,hwei,kfzhao,kfwong}@se.cuhk.edu.hk[1]
zywei@hlt.utdallas.edu[3], rolanchen@tencent.com[4]

Abstract. Microblogging websites have emerged to the center of information production and diffusion, on which people can get useful information from other users' microblog posts. In the era of Big Data, we are overwhelmed by the large amount of microblog posts. To make good use of these informative data, an effective search tool is required specialized for microblog posts. However, it is not trivial to do microblog search due to the following reasons: 1) microblog posts are noisy and time-sensitive rendering general information retrieval models ineffective. 2) Conventional IR models are not designed to consider microblog-specific features. In this paper, we propose to utilize learning to rank model for microblog search. We combine content-based, microblog-specific and temporal features into learning to rank models, which are found to model microblog posts effectively. To study the performance of learning to rank models, we evaluate our models using tweet data set provided by TERC 2011 and TREC 2012 microblogs track with the comparison of three state-of-the-art information retrieval baselines, vector space model, language model, BM25 model. Extensive experimental studies demonstrate the effectiveness of learning to rank models and the usefulness to integrate microblog-specific and temporal information for microblog search task.

Keywords: Microblogging Analysis, Online social network, Information Retrieval, Microblog Search, Experimental Study

1 Introduction

With the arrival of Web 2.0, microblogging websites, e.g. Twitter and Sina Weibo, now have become new and valuable source of information. As more and more

* This work is partially supported by General Research Fund of Hong Kong (417112), RGC Direct Grant (417613), and Huawei Noah's Ark Lab, Hong Kong. We would like to thank Junjie Hu, Prof. Michael R. Lyu and anonymous reviewers for the useful comments. This work was done when Zhongyu Wei and Junwen Chen were at The Chinese University of Hong Kong.

people join in these networks, the microblog messages they posted at different time cover a wide range of topics. Microblog posts are valuable due to the following reasons. 1) Information on microblogging websites rapidly updates, people can obtain latest information of many time-sensitive events, such as breaking news about missing flight MH370 or the latest information about iPhone 6. 2) There are rich of some small but useful tips in microblog posts, such as how to ace an interview in Google, and how to find delicious food in Hong Kong. These tips spread widely and quickly over social networks, which may be ignored by conventional social medias like webpages and newspapers. 3) Microblog posts are short and mostly in daily conversational style, thus are easy to read. People can get what they want with a quick scan.

Due to these overwhelming data, designing search tools to distinguish interesting and relevant microblog posts is crucial. To make good use of these information, we study the task of real-time and ad-hoc search microblog search. The goal is to retrieve "interesting" and "new" microblogging messages. A message is "interesting" means that it is relevant and covers content what users want. Also, the retrieved posts should be new, because people post millions of messages every day and out-of-date information values nothing. In particular, the users' information needs is represented by a query at a specific time, since they wish to see most recent and relevant information.

An intuitive approach is to directly apply conventional information retrieval models for microblog search. However, the characteristics of microblog posts challenge the current information retrieval models. Firstly, microblog posts are short and noisy. There are full of informal texts and unedited contents in microblog messages. Secondly, microblog posts are time-sensitive. Therefore, different from conventional search engines, microblog search engine should answer a query by providing a list of not only interesting but also newer relevant microblog posts. Thirdly, current information retrieval models are not designed for microblog specific features, such as emoticons, hashtags, urls, etc. These features are very useful in measuring relevance and would help to improve the performance of microblog retrieval model.

In this paper, we extend learning to rank framework to rank and retrieve microblog messages. Learning to rank is a combination of machine learning techniques and traditional ranking model. It makes prediction based on the difference of features. Specifically, we implement four state-of-the-art Learning to Rank Models, i.e. MART [4], RankBoost [3], Coordinate Ascent [10] and Lambda MART [15]. Also, we combine content-based, microblog-specific and temporal features into the learning to rank models.

To study the effectiveness of learning to rank models and proposed features, we evaluate the performance of our model using tweet set provided by TREC2011 and TREC 2012 microblog track. TREC 2011 data is for training and TREC 2012 is for test. In the experiment, we compare our learning to rank models with three baselines, i.e. cosine similarity, language model and BM25, all of which are state of the art conventional IR models.The experimental results confirm the effectiveness of learning to rank models for this task and demonstrates the

usefulness of microblog-specific temporal features in microblog retrieval. In particular, though we conduct experiment on Twitter dataset, it is trivial to utilize models and features proposed in this paper on other microblogging services, e.g. Sina Weibo and Tencent Weibo.

2 Related Work

Microblog retrieval has drawn tramendous attentions in recent years. Therefore, TREC introduced a track for ad-hoc microblog retrieval in 2011 [12, 13, 7]. Large tweet collections and annotations for various queries were released. Different approaches were investigated for microblog retrieval to overcome the special nature of microblog messages, e.g. short, noisy and time-sensitive characters of microblog posts. One of the main challenges in microblog retrieval is term mismatch due to short queries and short relevant messages, which renders conventional IR models ineffective. Researches tackled the term mismatch problem in microblog posts either by text clustering or query expansion. Another line of research in improving retrieval performance in microblog messages focused on using microblog-specific features to improve the performance of microblog retrieval [2, 9, 11, 5].

Duan et al. [2] used learning to rank models for tweet search. They studied some microblog-specific features, namely embedded URLs, mentions, hashtags, retweet behaviors, etc. They constructed a corpus for evaluation containing 20 self-selected queries and relevant tweets were crawled from Twitter Search[1]. Experiment results indicated the effectiveness of several features, namely, mention, URL and length. As the best system in TREC microblog 2011, Metzler et al. [9] also utilized learning to rank model incorporating 8 basic features.

Although researchers have proved the effectiveness of microblog message reranking for improving search accuracy, afore-mentioned works have the following problems: 1) the training sets used in some study were too small to produce a stable ranker [2]; 2) Features in some works were not rich enough[9, 11, 5]; 3) All of these works simply employed some single ranking model which were also query-insensitive.

Different from the previous, our work focused on improving the microblog search result by using the state-of-art learning to rank models instead of the conventional information retrieval approaches. Having large dataset from Twitter, we can extract various features and learn a stable ranker based on our multi-view features, which greatly improve the performance of our system.

3 Learning to Rank Model

Learning to Rank models, combination of machine learning techniques and conventional ranking model, attract increasing interests in the area of information

[1] `https://twitter.com/twittersearch`

retrieval [6]. In this paper, we extend a learning to rank framework for our real-time ad-hoc microblog search task. As a supervised learning method, learning to rank model can be formulated as a machine learning problem. A training instance, is a feature-label pair $< x, y >$, where x denotes a feature vector and y is the label corresponding to the ranking for microblog post t given query q. The essential issue in this step is how to extract feature vector x for $< t, q >$ pair.

Suppose that $F(\mathbf{x})$ is a function mapping a list of feature vectors $\mathbf{x}$ into a list of scores, and the loss function $L(F(\mathbf{x}, \mathbf{y})$ evaluates the difference between prediction result $F(\mathbf{x})$ and golden truth $\mathbf{y}$.The training process is to optimize $L(F(\mathbf{x}, \mathbf{y})$ and find optimal solution $F^*(\mathbf{x})$ with training data $< x_1, y_1 >, < x_2, y_2 >, ..., < x_n, y_n >$. So, after training, for any given query q, learning to rank models can automatically rank a corpus of microblog posts $T = \{t_1, t_2, ..., t_m\}$ according to the predictions of $F(\mathbf{x})$ and our search engine returns the top-k microblog posts.

Previous works propose various learning to rank models with different definition of prediction functions $F(\mathbf{x})$ and loss functions $L(F(\mathbf{x}, \mathbf{y})$, and proves them effective for many kinds of tasks. In this work, we utilize four state-of-the-art learning to rank models, i.e. MART [4], RankBoost [3], Coordinate Ascent [10] and Lambda MART [15].

Table 1 lists all the features we integrate to learning to rank models.

Table 1. Features Description

Feature Category	Feature Name	Feature Description
Content-based	CosSim	CosSim refers to the cosine similarity between t and q
	LM	Language Model denotes the uniform language model related to t and q
	BM25	BM25 represents the BM25 score of a microblog and a query
Microblog-Specific	#hashtags	#hashtags denotes the count of hashtags in t
	#match-hashtags	This is the number of words in q that also appear in t's hashtags.
	#words	This feature represents the number of words in t excluding stop words,
	Length	Length denotes #characters in t, including punctuations, emoticons, etc.
	#urls	#urls means the count of urls in t.
	#emoticons	#emoticons refers to the number of emoticons in t.
Temporal	gap	gap is the difference between q's query time and t's posting time.

4 EXPERIMENTS

We tested our learning to rank model on the dataset TREC microblog track, and evaluated the results from two evaluation metrics : precision@30 and mean average precision (MAP). Through comparison with traditional information retrieval models Vector Space Model, BM25 and Language Model, we demonstrated our learning to rank approach remarkably can outperform all these baseline models.

4.1 Data Collection and Set-up

Our dataset was provided by TREC 2011 and 2012 microblog track[2]. It consists of two parts: the microblog posts corpus for search, and the queries and golden sets of year 2011 and 2012 for evaluation. The raw microblog post corpus was crawled from the famous world-wide microblogging platform, Twitter. The queries and golden sets are available in TREC [3].

Table 2 summarizes the statistics of queries and golden set. There are 50 and 60 queries in 2011 and 2012, respectively. Every query contains one or multiple query keywords followed by a query time for ad-hoc search. For each query, the golden set annotates relevant or non-relevant information about 1,000 tweets in the corpus. Our learning to rank approaches used 2011 queries for learning while both our approaches and the baseline approaches were tested on the 2012 queries.

	QS 2011	QS 2012
# of queries	50	60
# of annotated tweets	40,855	73,073
# of highly relevant tweets	558	2,572
# of all relevant tweets	2,864	6,286

Table 2. The statistics of queries and golden set

The microblog track corpus contains tweets from Jan. 23rd to Feb. 8th, 2011, with 15,598,190 valid tweets. When a query is issued, only the relevant tweet posted before the query time will be returned as the result. According to the study of Wang et al.[14], considering the microblog posts after query time can in fact decrease the performance of microblog search system, because of microblog posts' time sensitive features.

To ensure the quality of microblog post retrieval from noisy and unstructured tweets, we preprocessed the raw tweet corpus as well as the queries via the following steps: 1) non-English tweets filtering; 2) tokenization and lemmatization; 3) case conversion (all letters were converted into lowercase); 4) Hashtag words duplication; 5) meaningless characters removing.

4.2 Experimental Results

We tested learning to rank models, with three state-of-the-art conventional IR models: Vector Space Model, Language Model and BM25, whose detailed information can be found in Manning et al. [8]. And following previous works relevant to BM25 models, we set $k_1 = 1.2$ and $b = 0.75$. The implementation of learning to rank models is based on RankLib [1] toolkit. The evaluation metrics were

[2] `http://trec.nist.gov/data/microblog.html`
[3] `https://github.com/lintool/twitter-tools/wiki/Tweets2011-Collection`

precision@30 and mean average precision (MAP for short), which are popular benchmarks in IR evaluation. Table 3 and Table 4 show the performance of base-line models and learning to rank models, respectively.

Table 3 indicates that conventional IR models performed badly and were not suitable for tweet retrieval. Among all the baselines, BM25 performed the best, because it can theoretically integrate the advantages of Vector Space Model and Language Model.

By comparing Table 4 to Table 3, we observed that learning to rank models remarkably outperformed all the baseline models. Specifically, the four learning to rank models achieved at least 32% improvement on precision@30 and 24.75% improvement on mean average precision. This experiment proves the effectiveness of learning to rank models on microblog search task.

Table 3. Performance of Conventional IR Models

	Vector Space Model	Language Model	BM25 Model
Precision@30	10.11%	3.84%	**21.92%**
MAP	6.64%	3.64%	**13.74%**

Table 4. Performance of Learning to Rank Models

	MART Model	Rank Boost	Coordinate Ascent	Lambda Mart
Precision@30	40.00%	33.00%	**40.56%**	35.65%
MAP	**32.03%**	25.75%	31.52%	28.33%

4.3 Feature Study

To illustrate the impact of content-based features, microblog specific features and temporal features, we conducted an analysis about different groups of features combined with learning to rank models. Figure 1 and Figure 2 illustrate the performance of learning to rank models combined with different combination of features, i.e. only content-based feature, only microblog specific feature, only temporal feature, content+microblog feature, content+temporal feature, microblog+temporal feature, and the combination of all three kinds of features, evaluated by precision@30 and MAP, respectively.

From the experimental result, we can learn the following results:

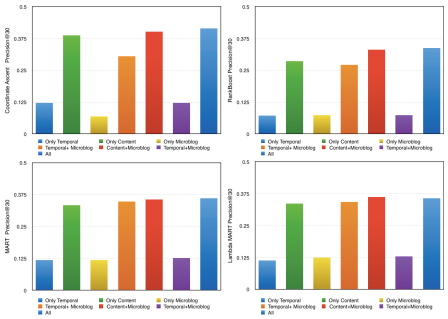

Fig. 1. Performance on Precision@30

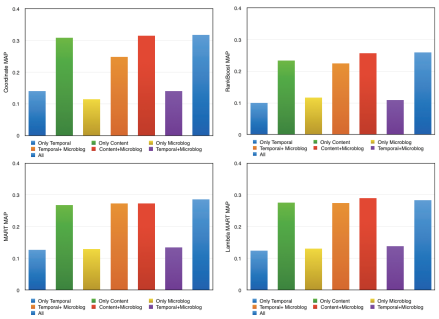

Fig. 2. Performance on MAP

– Different features have different influence on learning to rank models. Models are more sensitive to content-based features than that to temporal and microblog-specific features. Coordinate Ascent model and MART model had good performance with content-based feature. With only content based features and microblog specific features, the models can perform nearly as good as the models with all features. While with only temporal and microblog feature, models perform badly. However, the model combining all features had the best performance. So all features we proposed were effective and useful to learning to rank models.
– Different models have different sensitivity to different features. Among all models, Coordinate Ascent models and MART model are more sensitive to temporal and content-based features, while MART model and Lambda MART model are more sensitive to microblog-specific features.

In the practice, different queries may have different requirements on the features. For example, users may want to search microblog posts for some news about a topic. So the system should focus more on the temporal features. Coordinate ascent model should be a good choice. In some other case, users may want to get information from the online discussions. So the system is supposed to return microblog posts having relevant hashtags where Lambda MART model may be a good choice.

5 Conclusion and Future Work

In this paper, we address the importance of microblog search and point out the challenges of handling microblogging data. We find that conventional information retrieval models are not applicable for microblog search because: 1) Microblogging messages are short and noisy; 2) Conventional IR models cannot take full advantage of the microblog features into consideration. Through detailed analysis of microblogging data, we extract many useful features from content-based, microblog-specific and temporal aspects to model microblog messages, which are found to be useful. We deploy our models to retrieve tweets from

the tweet dataset provided by TREC 2011 and TREC 2012 microblogging track. The experimental results show that our models outperform all the baselines by doubling both precision@30 and mean average precision (MAP).

In the future, we will take the semantic information inside the query term and microblog posts content into consideration. Since microblog messages are short and concise, we should fully exploit the inherent meanings by designing microblog-specific PLSA (Probability Latent Semantic Analysis) or LDA (Latent Dirichlet Allocation) techniques. In addition, structure information, such as the social relationship network, can be used to improve the retrieval performance.

References

1. Dang, V.: Ranklib (2013)
2. Duan, Y., Jiang, L., Qin, T., Zhou, M., Shum, H.Y.: An empirical study on learning to rank of tweets. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 295–303. Association for Computational Linguistics (2010)
3. Freund, Y., Iyer, R., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. The Journal of machine learning research 4, 933–969 (2003)
4. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of Statistics pp. 1189–1232 (2001)
5. Han, Z., Li, X., Yang, M., Qi, H., Li, S., Zhao, T.: Hit at trec 2012 microblog track. In: Proceedings of the 21st Text REtrieval Conference (TREC 2012) (2012)
6. Hang, L.: A short introduction to learning to rank. IEICE TRANSACTIONS on Information and Systems 94(10), 1854–1862 (2011)
7. Lin, L., Efron, M.: Overview of the trec-2013 microblog track. In: Proceedings of the 23rd Text REtrieval Conference (TREC 2013) (2013)
8. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
9. Metzler, D., Cai, C.: Usc/isi at trec 2011: Microblog track. In: TREC (2011)
10. Metzler, D., Croft, W.B.: Linear feature-based models for information retrieval. Information Retrieval 10(3), 257–274 (2007)
11. Obukhovskaya, Z., Pervyshev, K., Styskin, A., Serdyukov, P.: Yandex at trec 2011 microblog track. In: Proceedings of the 20th Text REtrieval Conference (TREC 2011) (2011)
12. Ounis, I., Macdonald, C., Lin, J., Soboroff, I.: Overview of the trec-2011 microblog track. In: Proceeddings of the 20th Text REtrieval Conference (TREC 2011) (2011)
13. Soboroff, I., Ounis, I., Lin, J., Soboroff, I.: Overview of the trec-2012 microblog track. In: Proceedings of the 21st Text REtrieval Conference (TREC 2012) (2012)
14. Wang, Y., Lin, J.: The impact of future term statistics in real-time tweet search. In: Advances in Information Retrieval, pp. 567–572. Springer (2014)
15. Wu, Q., Burges, C.J., Svore, K.M., Gao, J.: Adapting boosting for information retrieval measures. Information Retrieval 13(3), 254–270 (2010)