The TREC Datasets in LETOR

Tao Qin, Tie-Yan Liu, and Hang Li

Microsoft Research Asia

tyliu@microsoft.com

The TREC datasets are also available in the LETOR package. The datasets contain features extracted from query-document pairs in the topic distillation task of TREC 2003 and TREC 2004, the corresponding relevance labels. They also include the evaluation results of several baseline ranking algorithms. In this document, we first introduce the topic distillation task of TREC, and then the features we extracted. After that, we describe the directories and formats of the datasets, as well as the baseline experimental results using the data.

1. Topic Distillation Task of TREC

The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U. S. Department of Defense, was started in 1992 as part of the TIPSTER Text program. Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. In particular, TREC has the following goals:

- to encourage research in information retrieval based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

Each TREC contains several tracks. For example, TREC 2004 contains the genomics track, the HARD track, the novelty track, the question answering track, the robust track, the terabyte track, and the web track. The goal of the web track is to investigate the retrieval behavior when the collection to be searched is in a large hyperlinked structure such as the World Wide Web. The web tracks in TREC 2003 and TREC 2004 used the .GOV collection, which is based on a January, 2002 crawl of .gov web sites. There are in total 1,053,110 html documents in this collection, together with 11,164,829 hyperlinks.

There are mainly three tasks in the web tracks: topic distillation, named page finding, and homepage finding tasks. Topic distillation task aims to find a list of entry points for good websites principally devoted to the topic. The focus is to return entry pages of good websites rather than the pages containing relevant

information themselves, since the entry pages provides a better overview of the coverage of a topic in the collection. TREC committee provides judgment for the topic distillation task. The human assessors make binary judgments to as whether a page is appropriate to a given query. That is, a page is judged relevant only if it is the entry page of some website which is principally devoted to the query topic. In this regard, this task is very similar to the Web search scenario. There are 50 queries and 75 queries in topic distillation tasks of the web track in TREC 2003 and 2004.

Many research papers [5][9][11][14] have been published which use the topic distillation task as their experimental platform. However, since the features and the data partitions are different in these papers, the corresponding experimental results are not directly comparable. To solve this problem, we try to extract standard features for each document-query pair in the topic distillation task and create data partitions in a standard way, such that future research on learning to rank can leverage the use of the datasets. We refer to them as "the TREC Datasets in LETOR", consisting of TD2003 and TD2004.

2. Features Extracted for the TREC Datasets

Since there are many documents (webpages) in the .GOV collection, when extracting features, we first used BM25 [12] to rank all the documents with respect to each given query, and then extracted features from the corresponding top 1000 documents. Note that some relevant documents (judged by the TREC committee) may not be in the top 1000 results of BM25. Therefore, in addition, we also extracted features from all the relevant documents for each query. As a result, in our datasets, some queries may have more than 1000 documents, while some may have less than 1000 documents (because they have less than 1000 returned documents).

When conducting feature extraction, we adopted the following principle:

(1) try to cover all the standard features in IR.

(2) try to reproduce the features proposed in the papers published in recent SIGIR conferences that also used the TREC data collection in the experiments.

With the principle, we extracted the following four categories of features. Note that, when extracting features, we conform to the original documents or papers. If the authors mentioned parameter tuning with regard to the feature, we also conducted tuning based on the whole dataset. If the authors only provide a default parameter and have not mentioned parameter tuning, we will use their default parameter directly in our feature extraction process.

1) Low-level Content Features

Low-level content features include term frequency (tf), inverse document frequency (idf), document length (dl) [1], and their combinations (e.g. tf*idf). For each of these features, we have four different values corresponding to the four fields of a webpage: body, anchor, title, and URL.

2) High-level Content Features

High-level content features include the outputs of BM25 [12] and LMIR [15] algorithms. We extracted four BM25 features in total, with the first one using the whole document for calculation, the second one using anchor text, the third one using title, and the last one using extracted title [5]. For LMIR, different smoothing methods (DIR, JM, and ABS) [15] and three fields (anchor, title, and extracted title) were used. As a result, there are nine language model features in total.

3) Hyperlink Features

Hyperlink features include PageRank [10], HITS [8] and their variations (HostRank [14], topical PageRank and topic HITS [9]). Since HITS and topical HITS have both authority and hub scores, there are 7 hyperlink features in total.

4) Hybrid Features

Hybrid features refer to those features containing both content and hyperlink information, including "hyperlink-based relevance propagation" and "sitemap-based relevance propagation" [11].

5) Feature List

In total, there are 44 features for each query-document pair. The complete feature list is shown as follows.

Table 1. All the features for the TREC datasets

Feature ID	Descriptions	References
1	BM25	[12]
2	dl of body	[1]
3	dl of anchor	
4	dl of title	
5	dl of URL	
6	HITS authority	[8]
7	HITS hub	
8	HostRank*	[14]
9	idf of body	[1]
10	idf of anchor	
11	idf of title	
12	idf of URL	
13	Sitemap based feature propagation*	[11]

14	PageRank	[10]
15	LMIR.ABS of anchor	[15]
16	BM25 of anchor	[12]
17	LMIR.DIR of anchor	[15]
18	LMIR.JM of anchor	
19	LMIR.ABS of extracted title*	
20	BM25 of extracted title*	[12]
21	LMIR.DIR of extracted title*	[15]
22	LMIR.JM of extracted title*	
23	LMIR.ABS of title	
24	BM25 of title	[12]
25	LMIR.DIR of title	[15]
26	LMIR.JM of title	
27	Sitemap based feature propagation*	[11]
28	tf of body	[1]
29	tf of anchor	
30	tf of title	
31	tf of URL	
32	tfidf of body	
33	tfidf of anchor	
34	tfidf of title	
35	tfidf of URL	
36	Topical PageRank*	[9]
37	Topical HITS authority*	
38	Topical HITS hub*	
39	Hyperlink base score propagation: weighted in-link*	[11][13]
40	Hyperlink base score propagation: weighted out-link*	
41	Hyperlink base score propagation: uniform out-link*	
42	Hyperlink base feature propagation: weighted in-link*	
43	Hyperlink base feature propagation: weighted out-link*	
44	Hyperlink base feature propagation: uniform out-link*	

Note that all the features followed by * are extracted based on recently-published SIGIR papers.

3. Files in the TREC Datasets

1) File Format

The label of each query-document pair is 0 or 1, where 0 stands for "not relevant", and 1 for "relevant".

We follow the format of SVM^{light} (http://svmlight.joachims.org/) input files to store the extracted features. Each line in the file represents a feature vector for a query-document pair, as shown below.

<label> <query id>:<value> <feature id>:<value> ... <feature id>:<value> # <info>

where <label> takes values from {0, 1, 2}, <query id> is an integer, <feature id> is as shown in Table 3, <value> is a float value of the corresponding feature, and document id is given at the end of each line as <info>.

An example line is as below,

2 qid:1 1:25.271132 2:58.000000 3:4.000000 4:5.000000 5:2.000000 6:0.000000 7:0.000000 8:0.000664 9:0.000022 10:0.000549 11:0.000489 12:0.001515 13:22.744019 14:0.000004 15:-9.333550 16:5.108520 17:-7.541900 18:-7.586890 19:-7.621820 20:1.110630 21:-7.222400 22:-7.156920 23:-7.071270 24:6.620330 25:-7.046290 26:-7.071270 27:22.744019 28:2.000000 29:1.000000 30:1.000000 31:0.000000 32:0.000042 33:0.000549 34:0.000489 35:0.000000 36:0.000000 37:0.000000 38:0.000000 39:0.000000 40:0.000000 41:0.000000 42:0.000000 43:0.000000 44:0.000000 #docid = 96

It means that for query id 1 and document id 96, the label is 2 (relevant). The 44 features extracted for the current query-document pair are (25.271132, 58.000000, 4.000000, …, 0.000000).

2) Directory and Data Partitioning

Here we take the TD2003 dataset for example, and the directory and data partitioning of the TD2004 dataset follow the same rule.

The feature file for the whole dataset of TD2003 is stored in the directory "TD2003\All". Furthermore, we partitioned the whole dataset into five subsets S1, S2, S2, S4 and S5, in order to conduct 5-fold cross validation. For each fold, we used three subsets for training, one subset for validation, and the remaining one for testing. The validation set is used to tune the parameters of ranking algorithms, such as the number of iterations in Neural Network training [2], the number of iterations in Boosting, and the coefficient parameter in the objective function of Support Vector Machines. In this way, we generated five datasets, which correspond to the directories "TD2003\Fold1", "TD2003\Fold2", "TD2003\Fold3",

"TD2003\Fold4" and "TD2003\Fold5" respectively.

Table 2. Data Partitioning for 5-fold Cross Validation

Sub Directories	Trainingset.txt	Validationset.txt	Testset.txt
Fold1	{S1, S2, S3}	S4	S5
Fold2	{S2, S3, S4}	S5	S1
Fold3	{S3, S4, S5}	S1	S2
Fold4	{S4, S5, S1}	S2	S3
Fold5	{S5, S1, S2}	S3	S4

We suggest that the users of the TREC Dataset conduct five-fold cross validation in their experiments.

4. Additional Information

Users of the TREC Dataset need to sign the license agreement as provided at the web site when they download the data. For any question or request regarding to this dataset, please send email to letor@microsoft.com.

5. References

[1] Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval. Addison Wesley, May 1999.

[2] Burges, C.J.C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G. Learning to rank using gradient descent, Proceedings of ICML 2005.

[3] Freund, Y., Iyer, R., Schapire, R., and Singer, Y., An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 2003 (4).

[4] Herbrich, R., Graepel, T., & Obermayer, K. Large margin rank boundaries for ordinal regression. Advances in Large Margin Classifiers, MIT Press, pp.115-132, 2000.

[5] Hu, Y. H., Xin, G. M., Song, R. H., Hu, G. P., Shi, S. M., Cao, Y. B., Li, H. Title extraction from bodies of HTML documents and its application to web page retrieval, Proceedings of SIGIR 2005.

[6] Jarvelin, K., and Kekalainen, J. IR evaluation methods for retrieving highly relevant documents. Proceedings of SIGIR 2000, pp.41-48, 2000.

[7] Jarvelin, K., and Kekalainen, J. Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems, 2002.

[8] Kleinberg, J. Authoritative sources in a hyperlinked environment. Journal of the ACM, Vol. 46, No. 5, 604-622, 1999.

[9] Nie, L., Davison, B. D., Qi, X., Topical link analysis for web search, Proceedings of SIGIR 2006.

[10] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank citation ranking: bringing order to the

Web, Technical report, Stanford University, 1998.

[11] Qin, T., Liu, T. Y., Zhang, X. D., Chen, Z., and Ma, W. Y. A study of relevance propagation for web search. Proceedings of SIGIR 2005.

[12] Robertson, S. E. Overview of the okapi projects, Journal of Documentation, Vol. 53, No. 1, 1997, pp. 3-7.

[13] Shakery, A., Zhai, C. X. Relevance Propagation for Topic Distillation UIUC TREC 2003 Web Track Experiments, Proceedings of TREC 2003.

[14] Xue, G. R., Yang, Q., Zeng, H. J., Yu, Y., and Chen, Z. Exploiting the hierarchical structure for link analysis. Proceedings of SIGIR 2005.

[15] Zhai, C. and Lafferty, J. A study of smoothing methods for language models applied to Ad Hoc information retrieval. Proceedings of SIGIR 2001, pp. 334-342, 2001.