Relevance Feedback by Exploring the Different Feedback Sources and Collection Structure

ZHANG Junlin SUN Le LV Yuanhua ZHANG Wei

Open System & Chinese Information Processing Center
Institute of Software, The Chinese Academy of Science
Email:junlin01@iscas.cn

Abstract: In HARD track of HARD 2005, we classify the 50 queries into 7 categories and make use of 3 kinds of feedback sources in various tasks. We find that the different kinds of queries perform differently in feedback tasks and the "CASE " and "EVENT" queries are more sensitive to the feedback source. We also explore the internal structure of corpus and try to estimate the distribution of relevant documents within sub-collections. The experiments show that this technology is partly effective and the main existing problem is how to predict the distribution more precisely.

1. Introduction

We participated in Hard track of HARD 2005 and our research mainly focuses on the following 3 aspects: 1. To classify all 50 hard queries into 7 categories to see whether the different kind of queries have various effects in feedback tasks; 2.Try to use various feedback sources to observe whether they perform equally in feedback tasks; 3.To explore the internal structure of corpus and try to estimate the distribution of relevant documents within sub-corpus according to the relevance feedback results.

We can draw the preliminary conclusions from the experimental results: 1.The different kinds of queries perform differently in feedback tasks and the "CASE " and "EVENT" queries are more sensitive to the feedback source. 2. The technology of exploring the internal structure of corpus in feedback task is partly effective and the main existing problem is how to predict the distribution more precisely.

In the following parts of the paper, we will describe our research goals and experimental results more clearly.

2. Research Goals

In HARD track of TREC 2005, we focus our research on the following 3 goals:

- **Query Category**

The queries of 2005 HARD track are the hardest queries selected from the previous TREC tests. As stressed by Cronen-Townsend et.al. [1], poorly-performing queries considerably hurt the effectiveness of an IR system. Many research work has been done to predict the difficulty of query [2,3,4,5,6]

So the first thing for us is to classify these hardest queries into several categories and analyze the reason why this type of query can't be satisfactorily processed by current IR technology.

The following table shows the query category we made in TREC 2005:

Category Name	Queries which belongs to the category	Feature patterns of query
CASE	307/325/353/374/378 /383/389/393/394/426/ 439/622/650/689	1. identify concrete cases of something… 2. identify individual or corp.which ….
EVENT	322/336/347/354/362/36 7 /408/448	1."identify instances of doing something…"
REASON	363/397/401/409/436 /639	1.what are the causes of something…
RELATIONSHIP	310/330/427/443	1.find document that discuss A and B…
MEASURE	341/344/435	1."what steps has been taken…"
STATUS	416	1.what is the status of something…
COMMON	303/314/345/372/375 /399/404/419/433/625 /638/648/651/658	Common questions

Table 1.Details of Query Category

Apparently the failure of the "CASE" and "EVENT" queries is due to the reason that information need is too general to find the relevant documents just under the current "bag of words" framework. As for the other type of queries, there is no apparent clue which can explain why the current IR technology fails to find good results. However, a very common problem by the TF.IDF approach [7,8](no matter what the category the query belongs to) is that the retrieved top documents always focus on just one theme even though the information need contains several themes. These irrelevant documents occupy most top positions. For example, the topic of query number 394 is "home schooling" while the top retrieved documents focus on just "home" or "schooling" instead of both of them. This indicates that the proximity information is a very important factor to improve IR performance under the TFIDF paradigm for many queries.

After classifying the query into several categories, we want to know weather the different type of queries will have different effect on IR performance in feedback tasks. The experimental results tell us that the query category really has different effect on the feedback performance.

- **Various Feedback Sources**

To evaluate the effect of various relevance feedback sources, we make use of 3

different collections as feedback source in our experiments: The corpus of TREC2005 HARD track (AQUAINT) , the corpus of TREC2004 HARD track(a collection of news from 2003 collated especially for HARD) and the web(using the Google to find out the relevant documents). We want to know answers of the following 2 questions:

1. Weather the different feedback source will bring different feedback effect?
2. Does the query category have effect on these various feedback source? If the answer is yes, what kind of effect it will be?

We firstly retrieval the different results from the above-mentioned 3 different corpus and extract the titles and 10 keywords of the top 15 initially retrieved document to form the CF for relevance judgment. Then the initial query is changed by adding the title and the keywords of all relevant documents into it to process the next retrieval (they are 3 different runs).

We call the AQUAINT corpus "inside corpus" and the other 2 corpuses as "outside corpus" in the following part of this paper for easier description.

- **Exploring the Collection Structure in Feedback Procedure**

The AQUAINT corpus consisted of three different newspapers: Xinhua news(XIE), New York times(NYT) and AWP. We suppose that the different news source may focus on different topics. For given query topic, the distribution of the relevant documents within these 3 sub-collections may different. For example, Xinhua news will have a bigger probability to publish the report about "three gorges project" than the other 2 news sources. So we plan to estimate this "relevant document distribution probability" through the feedback and hope this estimation parameter can be used to facilitate the IR performance. Here the probability can be regarded as the possibility that which sub-collection a relevant document should belongs to for any given query.

The first problem is how to estimate the parameters of the relevant document distribution within the 3 different news source given the query topic. We estimate the parameters as the following steps: Firstly, the relevant documents within top 15 search results (which are judged by NIST) are collected to form the "relevant doc set" for any given query topic. It's not hard to see which sub-set each document came from because the first 3 chars of document's name contain this information. For example, if the name of the document is "AWP200308122", we know that this document comes from AWP sub-collection. After labeling each relevant document with the news source, we can estimate the "relevant document distribution probability" as following :

$$P_i = \frac{R_i}{R} \qquad i = 1(XIE), 2(NYT), 3(AWP) \qquad (1)$$

where R_i means the number of relevant documents which belongs to different feedback source ; R means the number of all relevant documents in top 15 search results for any given query;

The second problem is how to tune the ranks of retrieved documents by applying the distribution parameters. We re-rank the initial retrieval result by the following formula:

$$FinalScore = \partial * OriginalScore + (1-\partial) * p_i \quad (2)$$

where ∂ =0.7 and p_i is the probability computed by formula 1.

3. Experimental Results and Analysis

3.1 Effect of Various Relevance Feedback Source

	Average precision	Change(compared with cassbase)	R-Precision	change(compared with cassbase)
cassbase	0.1514	null	0.2084	Null
cassgoogle	0.1342	-11.36%	0.2012	-3.45%
casstopdoc	0.1474	-2.64%	0.2054	-1.44%
cassself	0.2054	+35.67%	0.2554	+22.55%

Table 2. Performance of different feedback source

In order to observe the effect of various relevance feedback sources, we design 4 runs in our experiments :cassbase, cassgoogle, casstopdoc and cassself. Cassbase is a blind feedback run by adding the 40 keywords extracted from top 15 retrieved documents (from AQUAINT corpus) into the initial query and this run is regarded as the baseline. Cassgoogle is a run which use the WEB as feedback source and the title and 10 keywords of each relevant document (which are judged by NIST) are added into initial query to perform another retrieval. The casstopdoc and cassself are like the cassgoogle run except the different feedback sources. The casstopdoc run use the corpus of 2004 HARD track as feedback source and cassself run use the AQUAINT corpus(corpus of 2005 HARD track) as the feedback source.

The experimental results listed in table 2 show that the “outside corpus” as the feedback source decrease the IR performance as a whole compared with the baseline run while the “inside corpus” greatly increase the performance.

However, we analyze the performance of each query topic and found out that the query category effect the performance greatly. For easier description, we can compare the results of cassgoogle with cassself run. We found that the performance of most queries (20 queries among all 22 queries) from “CASE” and “EVENT” category in cassgoogle run decrease dramatically compared with the cassself run. However, for the other type of query, sometimes the cassgoogle win and sometimes cassself win. Among all 28 queries which don’t belong to “CASE” and “EVENT” category, 16 of them outperform the cassself in cassgoogle run. The bad performance of “CASE” and “EVENT” category query are main reasons to explain the failure of the “outside corpus” compared with.”inside corpus” as feedback source.

We can draw the following conclusions:

1. Using the “outside corpus” as the feedback source, the query category will be the main factor to decide whether the feedback source will help increase the IR performance. For most “CASE” and “EVENT” queries, it will decrease the IR performance if the “outside corpus” is used as the feedback source. While for other type of query, the effect of “outside corpus” as feedback source still need further research. That is, the “CASE”and “EVENT” query are much more sensitive to the feedback source compared with other type of query.
2. Compared with Blind feedback run, casstopdoc run decrease the performance slightly while cassgoogle decrease dramatically. We thought it’s maybe because the TREC corpus are all news paper and the google search result vary very much in the text format.

As for the reason why the “CASE” and “EVENT” query are more sensitive to the feedback source, we thought it’s maybe the relevant document of this type of query focus on concrete cases which involve many proper names or concrete event, So these documents share little information and the information from “outside corpus” will give little help to find relevant information in another corpus. While the “CASE” query will be effective to use the “inside corpus” as the feedback source because there are similar reports in the same corpus.

3.2 Exploring the Collection Structure

	NO Re-ranking	Re-ranking	Change
Group 1	(Cassbase) AP: 0.1514 RP: 0.2084	(Cassbasere) AP: 0.1479 RP: 0.2011	 AP:-2.31% RP:-3.5%
Group 2	(Cassallfb) AP: 0.1885 RP: 0.2463	(Cassallre) AP: 0.1799 RP: 0.2346	 AP:-4.56% RP:-4.75%
Group 3	(Cassallfb2) AP: 0.2044 RP: 0.2549	(Cassallfb2re) AP: 0.1921 RP: 0.2414	 AP:-6.01% RP:-5.29%
Group 4	(Cassself) AP: 0.2054 RP: 0.2554	(Cassselfre) AP: 0.1899 RP: 0.2418	 AP:-7.54% RP:-5.32%

Table 3. The details of re-ranking runs

We design 4 group experiments to test the re-ranking approach described in section 2 according to the formula 2**.** We can see from table 3 that the re-ranking technology through exploring the collection structure fails to increase the IR performance and with the increase of the initial retrieval performance, the negative effect has increased.

We also analyze each query to find out the reason why this technology fails. For convenient description, we take the cassself and cassselfre run as example(2 runs in group 4). It’s found that 20 queries benefit from this technology, 5 queries remain

unchanged scores and the other 25 queries suffer from the technology. As for the performance of other 3 groups, it's almost the same 20 queries benefit from the technology.

We thought whether the distribution parameter is precise or not have important effect on the performance. If the estimation is near the true distribution of the relevant document within sub-collections, the technology will help to increase the IR performance (sometimes dramatically change the performance) while it will has negative effect if the parameters estimation is far from the truth. So the conclusion is the technology is sometime effective and the problem of this technology is how to estimate the distribution more precisely.

4. Conclusion

In HARD track of HARD 2005, we classify the all 50 query into 7 categories and make use of 3 kind of different feedback source in various tasks. We find that the different kinds of queries perform differently in feedback tasks and the "CASE " and "EVENT" queries are more sensitive to the feedback source. We also explore the internal structure of corpus and try to estimate the distribution of relevant documents within sub-collections, the experiments show that this technology is partly effective and the main existing problem is how to predict the distribution more precisely.

ACKNOWLEDGMENTS

Partly supported by the National Natural Science Foundation of China under Grant No.60203007 and the new star plan of science & technology of Beijing under Grant No.H020820790130.

REFERENCES

[1] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query per-formance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299 -306, Tampere, Finland, 2002.

[2] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *Adcances in Information Retrieval, Proceedings of the 26th European Conference on IR Research,ECIR 2004*, pages 127-137, Sunderland UK, 2004.

[3] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. In *ACM Transactions on Information Systems (TOIS)*, volume 20(4), pages 357-389, 2002.

[4] Hu, X., Bandhakavi, S., and Zhai, C. (2003). Error analysis of difficult TREC topics. In *Proceedings of ACM SIGIR 2003* (poster)

[5] Voorhees, E. M. (2004). Overview of the TREC 2004 Robust Retrieval Track. TREC 2004 Conference Note Book,

[6] Terry Sullivan. Locating question difficulty through explorations in question space. In *Proceeding of the first ACM/IEEE-CS joint conference on Digital libraries*, pages 251–252.
ACM Press, 2001.

[7] G..Salton(1971).The SMART Retrieval System-Experiments in Automatic Document Processing. Englewood Cliffs,Prentice Hall,1971.

[8] Salton,G.(1983).Introduction to Modern Information Retrieval, McGraw-Hill