TREC12 HARD Track at ISCAS

Zeng Wu, Lin Du, Le Sun, Shiwei Ye*
Institute of Software, Chinese Academy of Science
Wuzeng01@iscas.cn; ldu@sonata.iscas.ac.cn
sunle@iscas.cn
Graduate School of Chinese Academy of Science.
Shwye@gscas.ac.cn

Abstract

Statistical model in retrieval has been shown to perform well empirically. Extended Boolean model has been widely used in business system for its easiness to be complemented and not bad results. In this paper, a statistical model and modified Boolean model and natural language processing techniques, shallow query understanding techniques are used and results show that even with very limited training corpus, an appropriate statistical model can greatly improve the performance. .

Keywords: TREC12, HARD, Boolean, statistical model, gradient decrease

1. Introduction

The HARD, which means high accuracy retrieval from documents, is a new track in TREC12. The goal of HARD is to achieve high accuracy retrieval from documents by leveraging additional information about researcher and/or the search context, through techniques such as passage retrieval, and using very targeted interaction with the searcher.

The key point of this track is to choose the most relevant text "granularity", the difficulty in choosing which may lead to that Ad-Hoc Track dwindled in the few last years, according to the purpose and genre metadata in the topic. The granularity may be a total document, a passage, or even a sentence. It is different from traditional full text retrieval. If one part, maybe a passage or several sentences, in a document best suit the topic, but the total score of this document is not high, from the point of HARD, this document should be the best document in the list. For every document in the collection, calculating the relevance between a topic and all the granularity parts individually is ideally. But it will cost so much computation. So the process is divided into two steps as Q/A system does.

The rest of the paper proceeds as follows. Section 2 outlines some background and related work. Section 3 introduces how to generate query words automatically. Section 4 explains the baseline run. Section 5 presents the final run. The evaluation is concluded in section 6. Section 7 is the conclusion and future work part.

2. Background and Related Work

The problem of HARD retrieval system design can be thought of as a problem in the combination of the following steps. The first step is to get key words from the topic, the second is to get relevant by using retrieval model and then to rank them, the

last is to locate topic information in the documents. This system is built totally by our site.

The purpose of previous TREC Ad Hoc track is to increase individual topic effectiveness. To generate query more accurately, some ideas can be learned from that community. In William. S. Cooper [2], they come up with such model like the following:

$$\log O(R \mid Q, D) \approx c_0 + \sum_{i=1}^{6} c_i X_i$$

They get coefficient by fitting the equation to the empirical data by means of a logistic regression analysis. (Hosmer & Lemesshow 1989) The statistical clues, X_i, are all based on the conventional frequency counts in query and document instead of using thesauri, parsing, phrase discovery, disambiguation, and other natural language processing or AI-like approaches. To the contrary, they felt it is a virtue of regression procedures explored here that they are more hospitable than most to the incorporation additional clues. However, an astute use of simple stem and document frequency information lift s one to a high plateau of effectiveness.

As mentioned above, search with information about searcher and search context can lift the results. So, the relevance feedback is also studied in this experiment. Relevance feedback, despite its long history in information retrieval research, has not been successfully adopted. The closest feature found in some search systems is “find more documents like this”. Query expansion techniques have been used in a number of systems to suggest additional search terms, with limited success. There are many reasons for the apparent failure of relevance feedback. The primary one is the difficulty of getting users to provide the relevance information. Simply providing “relevant” and “not relevant” button in the interface does not seem to provide enough incentive for user. For this reason, researchers are investigating techniques to infer relevance through passive measures such as time spent browsing page or number of links followed from the a page. Another reason is that identifying the correct context is not simple. Experiments [13] have shown that if a user can indicate relevant sections or even phrases in a document, relevance is more accurate.

To resolve these problems, the sophisticated interface design and good algorithm for inferring context are required.

In the second step, there is a lot of literature on approaches to information retrieval, we will not survey them all here. The focus, here, is on the modification of extended Boolean model and the statistical model. The modified Boolean model is used in baseline run.

The standard Boolean retrieval has following limitations

1) It gives counterintuitive results for certain types of queries.
2) It has no provision for ranking documents.
3) During the indexing process, it is necessary to decide whether a particular document is either relevant or non relevant with respect to a given index term.
4) It has no provision for assigning importance factors or weights to query terms.

The P-norm model is proposed as alternative to the Boolean model. P-norm

model has the ability to consider weighted query terms and provides a ranking of retrieved documents in order of decreasing relevance.

However, the statistical model look information retrieval as a problem in the combination of statistical clues. The design objective is to achieve as high a level of retrieval effectiveness as possible, consistent with reasonable theoretical and computational simplicity. In the final run, a statistical model is constructed. Then the compare between two models is evaluated.

The TREC Q/A track is designed to take a step closer to information retrieval rather than document retrieval. The researchers in this field mostly use statistical method. Abraham Ittycheriah applied Machine Translation ideas to the Q/A [3]. Because they have sufficient rules and weights, the answers are created from learning their known question and answer pairs in the open domain.

In getting the answer to a query, researchers usually use tagging or parsing tools to tag the query, then get the critical information including answer concepts, which are identified by categorizing queries using a method similar in spirit to extracting the named entities [8], the named focuses [9], and question-answer tokens [10]. Then the same procedure put on the Documents, and it will cost so much time. This process is always off line. Lastly, using different matching methods to generate the answer.

In retrieval, a query using the phrase such as “white house” is much more likely to be satisfied by a document using those two words in sequence than by one that has them separately. For some words may have distinctive meaning in the context of another word or in a larger phrase. This approach, however, requires that all sentences, whether in documents or in queries be segmented into phrases. This depends on the identity of the previous word generated. David R. H. Miller [5] bring forwarded three states Hidden Markov Model to identify two words phrase..

All the models mentioned above are built based on the statistical foundations which mean overwhelming majority of documents paired with relevant queries are available. In practices, it is usually difficult to come by.

As mentioned above, the HARD track has some familiarity with Ad Hoc track, Q/A track, and Interactive track. Some useful ideas and techniques can be learned from these tracks based on both HARD requirements and the resources available in hand.

3. Automatic Construct Key Terms

To take part in the HARD track, the system is built completely by us on the RedHat Linux platform. To make search on disk file more conveniently Berkeley DB-4.2 is introduced in the system.

In every topic, the sentence is generated from the <title> field, <descr> field and <narr> field. Then use Brill Tagger tool to tag it. In HARD topics, most sentences from <narr> field have such word like “on topic”, “off topic”. If no such phrases in the sentences, it will have negative words like “neither, nor, no”. In this system, the sentences having such words negative sentences is called negative sentences and words extracted from these sentences are called negative words. Others are called positive

sentences and words from these sentences are called positive words. After tagging these sentences, the words not tagged as "NNP ", "NN","NNS", "VBD","VBN" are abandoned, except the last word in sentences. For in examining the sample tagged files, the last word, a noun word from human, always is tagged as CD. Some words tagged as ADJ may in fact have some meanings, but limited to our resources, they can not be identified and be thrown off.

Now a word list named positive and a negative word list is constructed. In either list, every word is not a stop word and has been stemmed. It has a remark telling it from title field or <narr> field or <descry> field. It is clear that word from title field has more importance in retrieval. As for the same words in the same list or in the two lists, we also include it as if they were different words. The relevance between query q and document d can be calculated according to the following equation

$$w = positive_c * \sum w_+ - negative_c * \sum w_- + length_c * doc_length + location_c * \frac{doclength}{\sum word_location} +$$

$$qlentgth_c * query_length$$

1) w_+ is the weight of word in positive list; w_- is the weight of word in the negative list.

$$w_+ = (0.5 + \frac{0.5 * freq_q}{doclength}) * \log_2(\frac{N-n}{n}) \qquad (*)$$

$$w_- = (0.5 + \frac{0.5 * freq_q}{doclength}) * \log_2(\frac{N-n}{n})$$

$freq_q$ is the count of the word occurring in the document. doclength is count of all of the words in document d, n is count of this word occurring in all of the collection , N is the count of all of the document in the collection.

To get this equation, the equation is modified according to the document [2]. W_- or w_+ is the weight of w_i in negative list or in the positive list

2) word_location is the offset where the word occur in the document.

3) positive_c and negative_c, length_c, location_c, qlength_c are our statistical model parameters. But for the limited training corpus by hand, which is only the training topics provided by HARD, and the importance of these five parameters, the last three parameters are omitted.

4 Metadata and Clarification Form

4.1 CF

The clarification forms contains the following fields. The first field is composed of the title of the topic. The second field is composed of the words extracted from the sentences in the <descr> fields and on-topic words of the <narr> field. The third field is a list of negative words, which are extracted from off-topic section of the <narr> fields. If there is no negative word, this field is empty.

4.2 Metadata

For the tag RELATED-TEXT of the metadata, the relevant words extracted from the documents are added to the queries. If GENRE equals to ADMINSTRATIVE, the documents in HARDGOV corpus is returned. If is I-REACTION, the documents in the HARDGOV is not retrieved. For the metadata PURPOSE, if it is ANSWER, the simple method of Q/A is used. The following section is to do with the circumstance that the GRANULARITY is passage.

All other tags are not processed.

5.Baseline run

In our baseline run, assign positive_c = negative_c = 1;

For a certain topic, we rank the document according to the followings:

1) Calculate the weight according equation (1)
2) Ignore the document whose weight less than 0;
3) Rank the document according weight got in 1)
4) If the count of ranked documents is more than 1000, choose the first 1000 documents
5) For each document from ranked highest to ranked lowest, get the raw document content. For every word in positive list, the first location where the word occurred is the offset value shown in results file. The length is document length minus offset.

6.Final run

5.1 train positive_c, negative_c .

There is a statistical relation between the topic-document relevance and total positive word weights, negative word weights, words location, the context of words in relevant document, which can be used when the metadata granularity is passage. Because having involved the document length in get words weight, the normalization is not considered in this step. But there is only document and topic No in training relevance document and no other resources are available, so the model is simplified to two parameters as mention above.

In training relevance document, the document is remarked as 1 or 0.5 or 0, which display the document is hard-relevant, soft-relevant and non-relevant. So document remarked as 1 is more relevant than remarked as 0.5 and 0. It is the same with the document remarked with 0.5 and 0. Then to get such expressions like the following

For every certain topic

$$w_i > w_j \forall i, j$$

When document i is remarked as 1, as HARD marked, document j is remarked 0.5 or 0.

When document i is remarked as 0.5, as HARD marked, document j is remarked 0.

We can simply write W in equation 1 as

$$w_i = positive_c * w_{+i} - negative_c * w_{-i}$$

Because positive_c and negative_c is constant so the constant term is integrated in the two sides of equation, then

$$positive_c * w_{+} - negative_c * w_{-} > 0 \quad (2)$$

For one topic a list of such equations is got, and for total topics, it consist of a complete list of equations .Now a appropriate value for positive_c and negative_c is set to make expression (2) true in training corpus. We use Gradient Decrease Algorithm, commonly used in numerical calculation to get them.

5.2 locate the information

For all the words in the positive list and the negative list, the place of the first sentence which obtain the positive word is the offset value required in result file. The offset is the file length minus the offset.

5.3 work done especially for the request of some metadata

If GENRE equals to ADMINSTRATIVE, the documents in HARDGOV corpus is returned. If is I-REACTION, the documents in the HARDGOV is not retrieved.

For the tag RELATED-TEXT of the metadata, the words only already in the word list is extracted from the documents are added to the queries ignoring the fact that it has been in the list.

If the tag GRANULARITY is passage, the retrieval processing is composed of two stages: document retrieval and passage-level ranking. The document retrieval first gets all the relevant documents. Initially, the summary of a document is zero. From top passage to end one, if it contains a word in the positive list, the sum is added with one. If a passage contains a word in the negative list, the score is decreased with one. In the end, the sum of every passage is acquired. Then the maximum of them divide by the doc length is the new score of the doc. Then the doc list is ranked according to the new score.

7.Evaluation and Results

The Hard-rel judgment means that the document is relevant *and* it satisfies the appropriate metadata. The Soft-rel judgment means that document is relevant to the topic but that it does not satisfy the appropriate metadata. It either does not satisfy the PURPOSE, GENRE, or the FAMILIARITY items (the others are not document-level items).

In constructing the model, we do not count in the idf value in the topic, as the formula * show. For we think our model is a modified Boolean model and the word is noun in the sentences, which has a substantial meaning. The more they occur, the more important they are. And we have a desire to see what is happening without obeying classical theory. But from the tables, this thought does not accord with the fact.

There is a relation between the first score and last score in the retrieval, but we divide it subjectively.

In re-scoring the doc, there should be a similar expression with the expression 1. But

results are even worse when we manually check them. The reason is that the amount of the sample points is not sufficient. For the same reason, the parameter in form 1 is not precise enough. The model cannot satisfyingly predict the future.

The training corpus in our site is nothing but the corpus provided by the HARD, so to make parameter trained precise enough, only four topics are chosen for testing the results. It is not sufficient.

Document level retrieval results.

Table 1

Hard-rel criteria	Baseline run	Final run
Average precision	0.0324	0.0715
R-Precision	0.0679	0.1210

Table 2

Soft-rel criteria	Baseline run	Final run
Average precision	0.0368	0.0858
R-Precision	0.0702	0.1406

The following is an operational definition of passage recall and precision as used in the evaluation. For each relevant passage allocate a string representing all of the character positions contained within the relevant passage (i.e., a relevant passage of length 100 has a string of length 100 allocated). Each passage in the retrieved set marks those character positions in the relevant passages that it overlaps with. A character position can be marked at most once, regardless of how many different retrieved passages contain it. (Retrieved passages may overlap, but relevant passages do not overlap.) The passage recall is then defined as the average over all relevant passages of the fraction of the passage that is marked. The passage precision is defined as the total number of marked character positions divided by the total number of characters in the retrieved set. The F score is defined in the same way as for documents, assigning equal weight to recall and precision: F = (2*prec*recall)/(prec+recall) where F is defined to be 0 if prec+recall is 0. We included the F score because set-based recall and precision average extremely poorly but F averages well. R-precision also averages well.

In all of the above, a document is treated as a (potentially long) passage. That is, for topics where the granularity is "document" the relevant passage starts at the beginning of the document and is as long as the document. (These are represented in the judgment file as passages with -1 offset and -1 length, but are treated as described above.) For any topic, a retrieved document (i.e., where offset and length are -1) is again just a passage with offset 0 and length the length of the document.

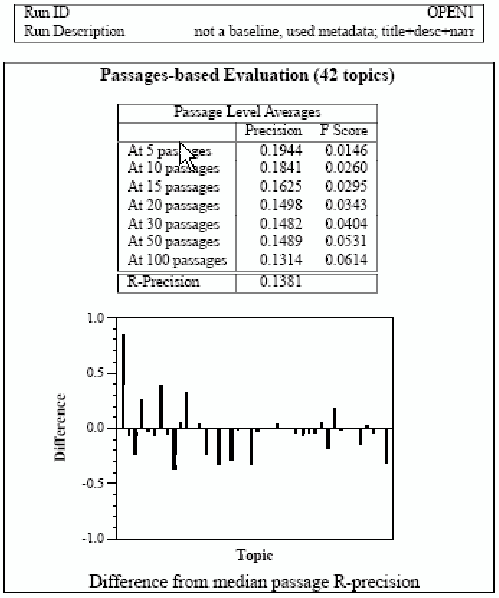
Using the above definition of passage recall, passage recall and standard document level recall are identical when both retrieved and relevant passages are whole documents. That is not true for this definition of passage precision. Passage precision will be greater when a shorter irrelevant document is retrieved as compared to when a longer irrelevant document is retrieved. This makes sense, but is different from standard document level precision.

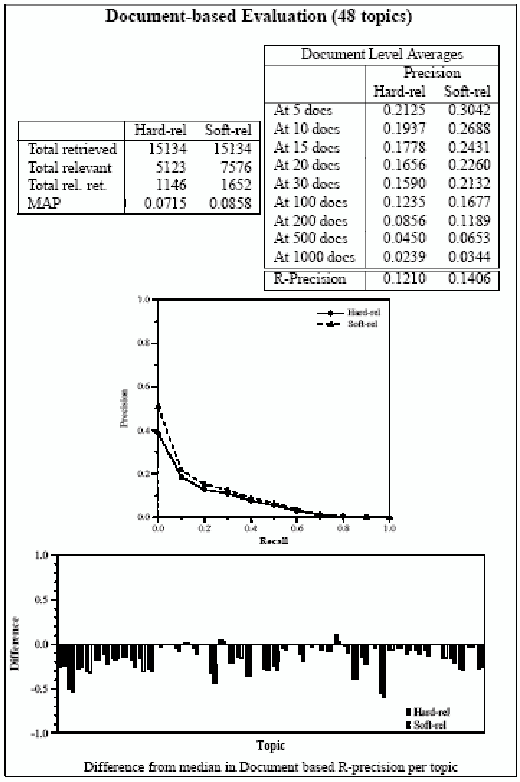
The following table is passage level results.

Table 3

	R-precision
OPEN	0.0954
OPEN1	0.1381

From the tables, the statistical elements partly overcome the model default in the baseline run.





8 Conclusion and Future work

The statistical model is effective, for using the same system, the latter results are twice better as much as the former.

The idf value is important whenever using any kind of retrieval model. It at least does not do any bad to the results.

Given more time and hands and more corpus, the equation (1) can be expanded in containing such elements as the doc length, query length, the word location occurring in the doc. And we will use Conjunctive Gradient Decrease or use MLP, using simulate anneal to relieve local minimum. From the contrast of the baseline run and final run, we are sure of performing better.

9.Acknowledge

This work is supported by China 863 Project(Grant No. 2001AA114040) and the National Science Fund of China under contact 60203007.

Reference

[1] E. Fox, S. Betrabet, M. Koushik "Extended Boolean Models" In "*Information Retrieval Data Structure & Algorithms*" pp393-418 1992

[2] William. S. Cooper, Aitao Chen, "TREC-3 working note: Experiments in the Probabilistic Retrieval of Full Text Documents"

[3] Abraham Ittycheriah, Salim Roukos, "TREC-11 working note: IBM's Statistical Question Answering System –TREC-11".

[4] Jinxi Xu, Ana Licuanan, Jonathan May, Scott Miller and Ralph Weischedel "TREC-11 working note: TREC 2002 QA at BBN: Answer Selection and Confidence Estimation"

[5] David R. H. Miller, Tim Leek, Richard M. Schwartz, "A Hidden Markov Model Information Retrieval System" In "*Annual ACM Conference on Research and Development in Information Retrieval Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval" B*erkeley, California, United States,pp214-221 1999

[6] Donna Harman "Ranking Algorithms" In "*Information Retrieval Data Structure & Algorithms*" pp363-392 1992

[7] Adam Berger, John Lafferty, "Information Retrieval as Statistical Translation" In *"Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval((SIGIR'99)",* pages 222--229, 1999

[8] Daniel Pack, Clifford Weinstein "The Use of Dynamic Segment Scoring for Language-Independent Question Answering" In *"HLT 2001 Presentations"*

[9] Adam Berger, Vibhu Mittal "Query-relevant summarization using FAQs" In *"Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL- http://citeseer.nj.nec.com/341776.html" 2000*

[11] Lynn Carlson, Daniel Marcu, Mary Ellen Okurowski. "Building a discourse-tagged corpus in the framework of rhetorical structure theory" In *Proceedings of the 2nd SIGDIAL workshop on discourse and dialogue, Eurospeech, Aalborg , Denmark*

[12] W. Bruce Croft, Stephen Cronen-Townsend, Victor Lavrenko "Relevance Feedback and Personalization: A Language Modeling Perspective" *In Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries, pages 49--54, 2001.*