Biomedical Test Collection with Multiple Query Representation

Borim Ryu
Seoul National University
103 Daehak-Ro, Jongno-Gu
Seoul, Korea
+82-2-766-3421
borim@snu.ac.kr

Jinwook Choi
Seoul National University
103 Daehak-Ro, Jongno-Gu
Seoul, Korea
+82-2-2072-3421
jinchoi@snu.ac.kr

ABSTRACT
The objective of this study is to validate pseudo gold standards using multiple queries for biomedical document collection. Aspect query is a quasi similar query of the original query text which has the same information needs. Basic idea is that single information need can be represented in various expressions. Four aspect queries per one original query were created manually by fifteen numbers of college students. By collecting the top ranked documents in retrieved sets generated by various queries, aspect query based pseudo gold standard is generated. In order to demonstrate its feasibility, we compared rank ordering based on human judgment and pseudo judgments. Experimental results showed the high correlations of 0.863 ($p<0.01$). Any difference among query worker group was observed and background knowledge about query topic did not seem to be prerequisite to create query sentences. Through the involvement of the experiment, we concluded that the method using multiple aspect queries can be a promising method for building relevance judgment without human experts.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software --- *performance evaluation.*

General Terms
Experimentation, Human Factors, Verification.

Keywords
Test collection, Evaluation of qrel sets, Correlation analysis, Gold standard

1. INTRODUCTION
In the medical area, almost all users want to retrieve biomedical bibliographic information. Several previous researches proved that the most common information resource used by clinicians was MEDLINE literature. However, it should be developed more sophisticated method of retrieving biomedical information due to its domain specific difficulty. In information retrieval (IR) point of view, test collection is used to evaluate system performance and to promote a common IR test bed for measuring effectiveness. A test collection is composed of documents, a set of queries and a list of all the collection documents that are relevant to each of query topics also known as qrels. Documents and queries are relatively easier to gather than creating relevance judgments, which requires significant effort and resources.

In this paper, we designed an experiment to validate a usability of pseudo-qrel generated by aspect queries for a biomedical test collection. Our study concerns a collection of highly domain-dependent biological and medical literature unlike many judgment-free approaches. In addition, we verified multiple query-based approaches with regards to many different factors, such as the number of top documents to be collected, query generating workers, etc.

This paper is organized as follows. Section 2 explains the detailed method about aspect query considered in this study and experimental system chosen for the experiment. Section 3 shows our experimental results and discusses correlations on different gold standards, between the original relevance judgment qrel made by TREC and aspect qrels found by experimental results and focuses on the performance of aspect qrel generated from aspect queries. Our conclusions are presented in section 4.

2. PAST WORK
Evaluation has always been focused as a crucial component of information retrieval. A test collection is used to evaluate the search algorithm or system performance. It consists of three main factors: document set, query topic of user's information need and relevance judgments of which documents are pertinent to each query, also called as qrels. It is clear that creating relevance judgment is a labor intensive task of human experts. Due to huge amount of time and effort, or even financial support, a variety of recent studies aimed to develop the way of building test collection with very few or even none relevance judgments.

An annual information retrieval conference and competition (TREC) is to support and promote further research in information retrieval society. Every year, TREC builds and distributes text collection; approximately 50 to 100 research groups submit their runs, which contain 1,000 numbers of best matching documents against TREC topics. The union of top documents from each run (referred as system pool in terms of TREC) is manually assessed by human experts for the exact relevance. TREC aims diverse searches through the assumption that each research group used their own searching system. As follows, the pooling method is widely applied as a means of sorting out relevant documents

within a large test collection. For each query, merging the retrieved output from diverse search system formed a pool. In this way, the assumption that nearly all relevant documents could be found in the pool is based. A top of the document pool is manually assessed by the experts for relevance, thereby forming the query-relevance document, qrel set.

In previous studies, Salton, Fox, and Wu (1983) described that seven different means of processing the query were run on a retrieval system, the documents retrieved by each means were merged (duplicates were removed) and the resulting pool examined by relevance assessors. In fact, Soboroff (2001) hypothesized that statistical sampling from documents in a system pool could generate a set of "pseudo-qrels". Efron (2009) suggested the concept of "query aspects" to construct pseudo-qrels without human assessment. The notion of aspect query is a textual instantiation of user's information need. The study aimed to generate new query which can reflect differing aspects of the original topic and make pseudo-qrels on behalf of human assessment.

However, it was not proven and characterized whether this method could be chosen to create gold standard qrels in a specific domain. In particular, biomedical literature retrieval should be able to handle domain dependent terms. There are few kinds of developed test collection to evaluate the performance of biomedical documents retrieval. We focused documents which topic is relevant to biology and medical science in this study.

3. METHOD

To verify pseudo-gold standard for a collection without human assessment, we firstly made multiple aspect queries regarding fifty topic in 2004 TREC genomics track. Fifteen college students read original topic and made four similar queries per one topic. We ran each query against a subset of MEDLINE document set along with BM25 probabilistic retrieval model in Terrier IR test beds. Based on the assumption that collecting top 100 documents retrieved by aspect query are possible to be relevant, aspect-qrel was generated. For the convenience we named the pseudo-relevance judgment with aspect queries as aspect-qrel. Figure 1 illustrates the overall experimental method.

3.1 Dataset

We utilized TREC Genomics track data as a test collection. In 2004 Genomics track, it contains 4,591,008 MEDLINE references on biology, medical or pharmacology and 50 topics derived from interviews eliciting information needs of real biologists. The topics are formatted in XML format and have the following fields: ID, TITLE, NEED, and CONTEXT. ID is the number of topic and TITLE is an abbreviated statement of user information need, usually used as search query. The element in NEED field is a full statement of original information need. In CONTEXT field, background information to place information need is described. Table 1 shows the collection data used in this study. Relevance judgments corresponding to each query are provided using the scale of 'definitely relevant', 'possibly relevant', and 'not relevant'. In our experiments, we limited relevant documents to those judged as definitely relevant; thus, documents obtained by aspect queries assumed to be definitely relevant.

Table 1. Dataset used for the experiment

Corpus	# Documents	# Topics	# Systems
2004 TREC Genomics track	4,591,008 MEDLINE Documents	50	46 runs

3.2 Multiple aspect queries

Aspect queries of the original query topic were generated manually by fifteen recruited college students. These new queries were created with regards to 2004 Genomics track topics, such as correlation between DNA repair pathways and skin cancer or substrate modification by ubiquitin, or cause of scleroderma. Every participant was given a set of 50 topics, and read all fields in the topic. The topic was consisted of topic ID, TITLE, NEED, and CONTEXT. A short title usually delivered as the type of query, along with an abbreviated statement of information need as usual that might be submitted to the IR system.

The number of aspect queries to construct pseudo-qrels with top retrieved documents was set to 4. We assumed that there should be a transition or differential threshold as the number of aspect queries increasing. However, there was any regular formula within the number of queries.

Aspect query is the elaboration, specification, or paraphrase of the original topic. During the aspect query creation, new query was made against the original one in order to express different facets. All participants could refer given topic fields and other online resources such as Korean Medical Library Engine (KMLE) and MeSH by NLM websites. We obtained four aspect queries for each of 50 Genomics track topics. For example, topic number 8 'correlation between DNA repair and skin cancer' had four aspect queries as follows: DNA repair gene mutation in skin cancer, DNA repair and skin carcinogenesis, pyrimidine dimer removal and XRCC3 associated with melanoma skin cancer.

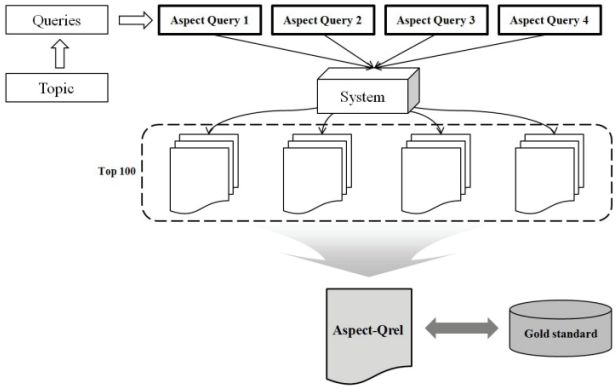

Figure 1. Overall process in this study

3.3 Pseudo gold standards

To create a set of pseudo-qrels with aspect queries for a given topic, we ran each aspect queries against the MEDLINE document collection using BM25 probabilistic retrieval model. BM25 is considered to be a seed system in this approach. This model was implemented on Terrier IR platform.

During the run 1000 documents were retrieved against each aspect queries. Aspect-qrels were made by summing the union of top K documents retrieved for all four aspect queries. Duplicated documents were removed. The tunable parameter K is set to 100 in this process. If there are too many duplicated documents in retrieved pool, the number of collected documents considered to be relevant to the query became smaller.

Search result with the original TREC topic and documents was evaluated according to the genuine qrel by human assessment and the aspect-qrels by multiple aspect queries. Retrieval evaluation measure is described in next subsection.

3.4 Evaluation measure

In order to evaluate our experimental results, we primarily used mean average precision (MAP) as the evaluation metric for retrieval effectiveness. MAP is the mean value of the average precisions computed from multiple queries where the average precision of each query is calculated by the average on precisions at each retrieved relevant document. MAP is frequently used as a good measure of the overall ranking accuracy. In all experiments, the measures were evaluated for the 100 top-ranked retrieved documents. We classified MAP according to type of gold standard used to compute, qrels made by TREC and aspect-qrels generated in this study. In particular, we designated MAP value computed by aspect-qrels as aMAP.

4. EXPERIMENTS

The correlation between aspect-qrels and those of TREC judged by human expert was evaluated.

In 2004 TREC Genomics track, forty-six kinds of runs with different retrieval scheme were submitted. With these retrieved results, we calculated the mean value of the average precisions with each qrels. We formulated rank orderings for 46 IR runs according to priority of MAP and aMAP in descending order. We examined rank orderings based on different qrels so as to analyze how correlates they are.

In order to analyze correlations, we measured rank correlations with two metrics, Kendall's tau and Spearman's rho coefficient. If the rank ordering by aMAP correlates highly with the ranking by MAP, it is likely to say that aspect query based qrels without human assessment could be confirmed the validity for a gold standard as well as human-judged qrels is worth. The significance test was performed under p-value <0.01.

4.1 Ranking correlations

In general, Spearman's correlation coefficient (Rho) is widely used for measuring relations among rank orderings. In our experimental result, rho coefficient showed high correlations by up to 0.863 between rankings derived by TREC qrels and aspect query based qrels (Table 2). The average value of whole correlations were 0.6694 in tau and 0.831 in rho under p-value<0.01. Table 2 shows tau and rho rank correlations between system rankings obtained by the official TREC MAP and aspect query based on aMAP.

Table 2. Rank Correlations with MAP computed by TREC qrel and aMAP computed by aspect query based 15 qrels. Each correlation was calculated under p < 0.01. The highest correlation value in bold.

Aspect query based qrel types	Kendall's Tau	Spearman's Rho
Bio 1	0.674	0.837
Bio 2	0.707	0.858
Bio 3	0.651	0.862
Bio 4	0.664	0.821
Bio 5	0.681	0.83
Med 1	0.632	0.791
Med 2	0.647	0.813
Med 3	**0.710**	**0.863**
Med 4	0.648	0.819
Med 5	0.670	0.833
Tech 1	0.666	0.824
Tech 2	**0.710**	0.827
Tech 3	0.673	0.861
Tech 4	0.661	0.83
Tech 5	0.69	0.848

4.2 Applications

4.2.1 Seed system

In our approach, BM25 retrieval model was applied as our "seed system" to retrieve queries and document. Our seed system used porter's stemmer and stop word removal. One might oppose that the correlations we have observed in this study seemed to be unclear because the BM25 probabilistic retrieval model was known to perform well. We tested other retrieval models and compared system performance with BM25. Hiemstra's language model and PL2 model based on Poisson estimation for randomness theory were concerned as variance (Table 3) The result of this process led to further considerations of choosing retrieval seed system.

Table 3. Baseline correlations used different seed systems: BM25, Language model and PL2. Each correlation was calculated under p < 0.01. The highest correlation value in bold

	BM25	Hiemstra's Language model	PL2
Kendall's tau	0.678	**0.688**	0.645
Spearman's rho	0.849	**0.86**	0.821

4.2.2 Subgroup

Aspect queries were generated manually in this experiment. 15 recruited college students had many different major and those were biology related, medical related and engineering field. We categorized the aspect query creators depending on their major background: Biology related group 1, medical and public health related group 2 and engineering related group 3. Participants were classified into three categories: Bio, Med and Tech group. There were 5 students in each group. In the Bio group and Med group, all students had an experience on studying life science. All students were able to understand the topic and fluent in English. Their task is to rewrite the query according to the original query and information need. They could refer a variety of online resources (Google, Korean medical terminology library site, MeSH, etc). Relevance judgment for the query and document is not necessary for the experiment.

Our intuitive assumption was that there should be differences on correlation if the aspect query creator had background knowledge on genetics or biochemistry. However, there was no marked difference among the groups unlike our hypothesis.

Unlike our intuitive assumption, the correlations among three groups were quite similar. Although no difference in correlations was shown, an attempt to in-depth analysis based on background knowledge was meaningful try in a sense. Figure 2 shows the additional analysis on worker groups. Each participant's correlations on Kendall's tau and Spearman's rho values were

illustrated in horizontal line. The arithmetical mean of rho coefficient among groups was 0.8413, 0.8237 and 0.8425. Surprisingly, medical related group was shown lowest correlations. This might be due to the lack of contributions in creating queries.

Through this study, making relevance judgment using aspect queries without human expert assessment turned out to be a possible method to build a test collection. The field specificity considered to be a barrier for information retrieval research in biomedical domain became lower to some extent. Hence, the method using multiple aspect queries to create a gold standard seems to be applied without biology background qualifications. It is expected that biomedical information retrieval study could be promoted in the future.

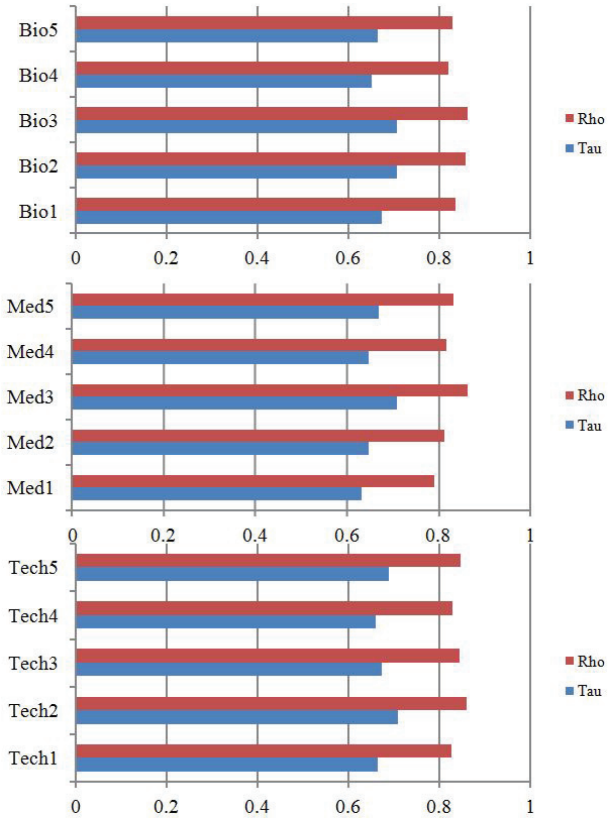

Figure 2. Additional analysis result on worker group's background knowledge

5. CONCLUSIONS

Aspect query, as the elaboration, specification, or paraphrase of the original topic, was generated manually and focused to build query-document relevance set by combination of top-ranked documents from the retrieved result. In our experiment, rankings assessed by aspect query based qrels correlates highly with ranking based on human relevance judgments. The correlations of rank orderings by judgment-free aspect qrels and by human-assessed qrel show a quite high correlations by up to 0.863, average value of 0.831 ($p<0.01$). As a result, we could verify creating relevance judgments without human experts by combining the top rank documents retrieved by a number of multiple queries. It is expected to contribute medical information retrieval and evaluation study.

6. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2010-0028079, 2011-0018259).

7. REFERENCES

[1] Efron M. *Using multiple query aspects to build test collections without human relevance judgments*. Advances in Information Retrieval 2009:276-87.

[2] Hersh, W., et al. *OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research*. in *the 17th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. 1994. Dublin.

[3] Sanderson M., Joho, H. *Forming test collections with no system pooling*. in *the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. 2004. Sheffield.

[4] Soboroff, I., Nicholas, C., Cahan, P. *Ranking retrieval systems without relevance judgments*. in *the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. 2001. New Orleans.

[5] Wu S, Crestani F. *Methods for Ranking Information Retrieval Systems without Relevance Judgments. in the 2003 ACM Symposium on Applied Computing*. 2003. Melbourne.

[6] Voorhees EM. *Building a Question Answering Test Collection*. in *the 23rd annual international ACM SIGIR conference on Research and Development in Information Retrieval*. 2000. Athens.

[7] Sanderson, M., and Braschler, M., *Best Practices for Test Collection Creation and Information Retrieval System Evaluation*, TrebleCLEF technical report, ISBN: 978-88-88506-93-7, 2009.

[8] Heppin KF, *MedEval: a Swedish medical test collection with doctors and patients user groups*. Association for Computational Linguistics 2010.

[9] Aslam A, Savell R, *On the effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments*. in *the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. 2003. Toronto.