Original Article

Healthc Inform Res. 2012 March;18(1):65-73.
http://dx.doi.org/10.4258/hir.2012.18.1.65
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

An Evaluation of Multiple Query Representations for the Relevance Judgments used to Build a Biomedical Test Collection

Borim Ryu, BS, Jinwook Choi, MD, PhD
Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul, Korea

Objectives: The purpose of this study is to validate a method that uses multiple queries to create a set of relevance judgments used to indicate which documents are pertinent to each query when forming a biomedical test collection. **Methods:** The aspect query is the major concept of this research; it can represent every aspect of the original query with the same informational need. Manually generated aspect queries created by 15 recruited participants where run using the BM25 retrieval model in order to create aspect query based relevance sets (QRELS). In order to demonstrate the feasibility of these QRELSs, The results from a 2004 genomics track run supported by the National Institute of Standards and Technology (NIST) were used to compute the mean average precision (MAP) based on Text Retrieval Conference (TREC) QRELSs and aspect-QRELSs. The rank correlation was calculated using both Kendall's and Spearman's rank correlation methods. **Results:** We experimentally verified the utility of the aspect query method by combining the top ranked documents retrieved by a number of multiple queries which ranked the order of the information. The retrieval system correlated highly with rankings based on human relevance judgments. **Conclusions:** Substantial results were shown with high correlations of up to 0.863 (p < 0.01) between the judgment-free gold standard based on the aspect queries and the human-judged gold standard supported by NIST. The results also demonstrate that the aspect query method can contribute in building test collections used for medical literature retrieval.

Keywords: Information Storage and Retrieval, Evaluation Studies, MEDLINE, Correlation Studies, Gold Standard

Submitted: January 12, 2012
Revised: February 21, 2012
Accepted: March 10, 2012

Corresponding Author
Jinwook Choi, MD, PhD
Department of Biomedical Engineering, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 110-799, Korea. Tel: +82-2-2072-3421, Fax: +82-2-745-7870, E-mail: jinchoi@snu.ac.kr

I. 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.

Evaluation has always been a critical component of information retrieval (IR) [1]. A number of information retrieval studies have focused on evaluating system performance with controlled experimental setting, so-called a test collection [2-5].

A test collection has three basic components which are document set, query topic of user's information need and relevance judgments. A set of relevance judgment of which documents are relevant to each query is also called query based relevance sets (qrels) and it is indispensable to quantify how accurate the system performs. However, it is the most difficult part to be constructed as it requires tremendously labor intensive task by human experts. Fundamentally, this process is done by several person who make a decision of which document is pertinent to the query topic after reading the entire documents and queries. Due to huge amount of time, effort, and even financial support, a variety of recent approaches to create test collections with very few or no relevance judgment have been previously studied [6-10].

Soboroff et al. [8] hypothesized the statistic sampling from document pool could generate a set of "pseudo-qrels". Wu and Crestani [9] had an advanced work regarding judgment-free evaluation by concerning the relationship among the retrieved documents in a pool. Efron [6] suggested an approach that utilize "aspect query" to construct pseudo-qrels without human assessment. The notion of aspect query is a textual instantiation of user's information need. It can be related in diverse ways, such as paraphrasing, generalizing or emphasizing of the information need. Unlike previous approaches, he aimed to generate new query which can bring up differing facets of the topic and make pseudo-qrels on behalf of human assessment. However, it was not proven whether this method could be chosen to create gold standard qrels in biomedical domain. In addition, it was also needed to be analyzed such as worker's distribution in query generation or the number of queries enough to take out appropriate relevant document in detail.

In the medical domain, the medical terminology and particular background knowledge is highly specific than any other one. Furthermore, biomedical information retrieval should be able to handle those domain related terms [11,12]. Because of such specificity, there are few kinds of developed test collection to evaluate the performance of biomedical documents retrieval system [4,5].

We designed an experiment to validate a usability of whether the method using aspect query could be adopted to build a biomedical test collection. The core idea of this paper is about the unique aspect of a query. For clarity, we explain the concept of aspect query. It might be an elaboration, rephrasing, specification, or generalization of the original query. Data used in this study was supported by National Institute of Standards and Technology (NIST), the main organization of Text Retrieval Conference (TREC) [13]. TREC Genomics track - one of the subtasks in TREC track, aims to encourage biomedical retrieval research [14-16]. Genomics collection contains subset of MEDLINE document sets and biological or clinical query topics and it was used to generate new aspect query. The submitted result runs at 2004 Genomics track were used to compare rank orderings according to the different gold standard.

The objective of this study is to examine the method using multiple query representation to build qrels regarding biomedical literature. In this paper, we describe an experimental evaluation that was analyzed rank orderings by a number of multiple aspect queries based qrels from single IR system. Our study is appealing to two main senses. First, unlike most judgment-free evaluation research, a collection of highly domain-dependent biomedical literature is concerned. Second, we verified aspect query-based approach 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. Section 4 discusses correlations on different gold standards, between the original qrels 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 also presented in section 4.

II. Methods

To verify aspect query method which used to generate gold standard for a collection without undue human assessment, we firstly made multiple aspect queries regarding fifty kinds of topic in TREC genomics track. We ran each query against a subset of MEDLINE document set along with BM25 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. Figure 1 illustrates the methodology used in this experiment.

In this study, forty-six of submitted result data at 2004 TREC genomics track were supported by NIST and used to compare for validation of proposed method [17]. We evaluated the retrieval performance with different gold standards, made by TREC human assessors and by aspect queries. According to the mean average precision (MAP) score, rank ordering was concerned to be calculated the correlations among the qrels through this study.

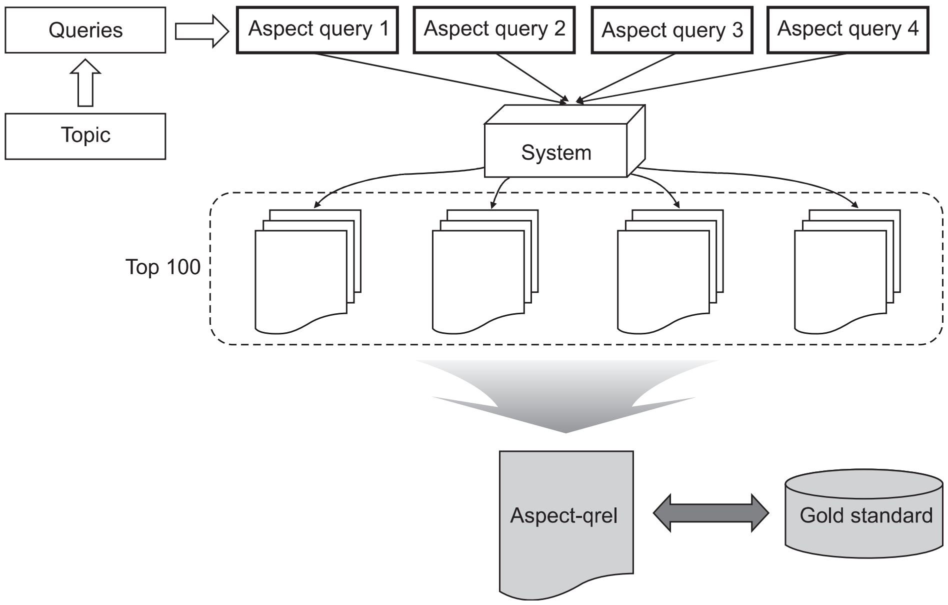

Figure 1. Block diagram of experimental methodology.

Table 1. Data sets used for experimentation

Corpus	Documents	Topics	Systems
TREC 2004 genomics track data	4,591,008 MEDLINE documents	50	46 runs

1. Test Collection

We used a collection from TREC genomics track as a test collection. The test collection is a subset of the MEDLINE database, which is itself a bibliographic database of important, peer-reviewed medical literature maintained by the National Library of Medicine (NLM). In 2004 genomics track, it contains 4,591,008 MEDLINE references on biology, medical or pharmacology and 50 topics (queries) derived from interviews eliciting information needs of real biologists [18]. Table 1 shows the collection data used in this study. Each query consists of information needs for which a biologist might search the literature and that retrieve reasonable documents. 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.

2. Creating Multiple Aspect Queries

A strong focus in information retrieval task is to discover pertinent items, which could be relevant to user’s information need. To accomplish this objective, aspect queries were generated manually by 15 recruited college students.

Aspect queries were created with regards to biomedical topics, such as correlation between DNA repair pathways and skin cancer, substrate modification by ubiquitin, or cause of scleroderma. All of participants were given a set of 50 topics, and they read all fields in the topic. The components of topic structure were topic id, title, need, and context. A short title could be viewed as the type of query, along with an abbreviated statement of information need as usual that might be submitted to the IR system. In need field, a full statement of information need is described. Context field was written about background information to place information need in context.

During the aspect 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 Medical Subject Heading (MeSH) by NLM websites [19,20]. Finally, 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.

3. Creating Gold Standards Using Aspect Queries

To create a set of pseudo-qrels with aspect queries for a given topic, we ran each aspect queries against the document collection, using BM25 retrieval model. BM25 is considered to be a seed system in our approach. This model was implemented Terrier search system, an open-source IR platform developed by University of Glasgow.

During the run 1,000 documents were retrieved against each aspect queries. Aspect-qrels were attained by collecting the union of top K documents retrieved for all four aspect queries. Duplicated documents were considered as one and others were removed. The tunable parameter K is set to 100 in this process (Figure 2).

4. Evaluation Measure

In order to evaluate our experimental results, we primarily used 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 serves as a good measure of the overall ranking accuracy, and it favors systems that retrieve relevant documents early in the ranking. In all experiments, the measures were evaluated for the 100 top-ranked retrieved documents. We classified MAP according to gold standard used to compute, qrels made by TREC and aspect-qrels generated in this study. In particular, MAP computed by aspect-qrels was called aMAP.

5. Rank Correlation Analysis

Correlation analysis is the method to compute the statistical relationship between two quantities. Furthermore, a rank correlation coefficient is used to measure the degree of similarity between two rankings and to evaluate its significance, such as Spearman's rank correlation coefficient (rho) and Kendall's rank correlation coefficient (tau). The coefficient is inside the interval -1 to 1 and assumes if the value is -1 then one ranking is the reverse of the other, and if the value is 1 then two rankings are identical. In general, Spearman's correlation coefficient is widely used for measuring relations among rank orderings.

The question to figure out in our analysis is, to which degree does evaluation performance calculated by judgment-free aspect-qrels correlate with evaluation performance calculated by original human-judged TREC qrels. In our case, evaluation performance by aspect-qrels is aMAP and the other one is MAP. 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 much as human-judged qrels is worth. The significance test was

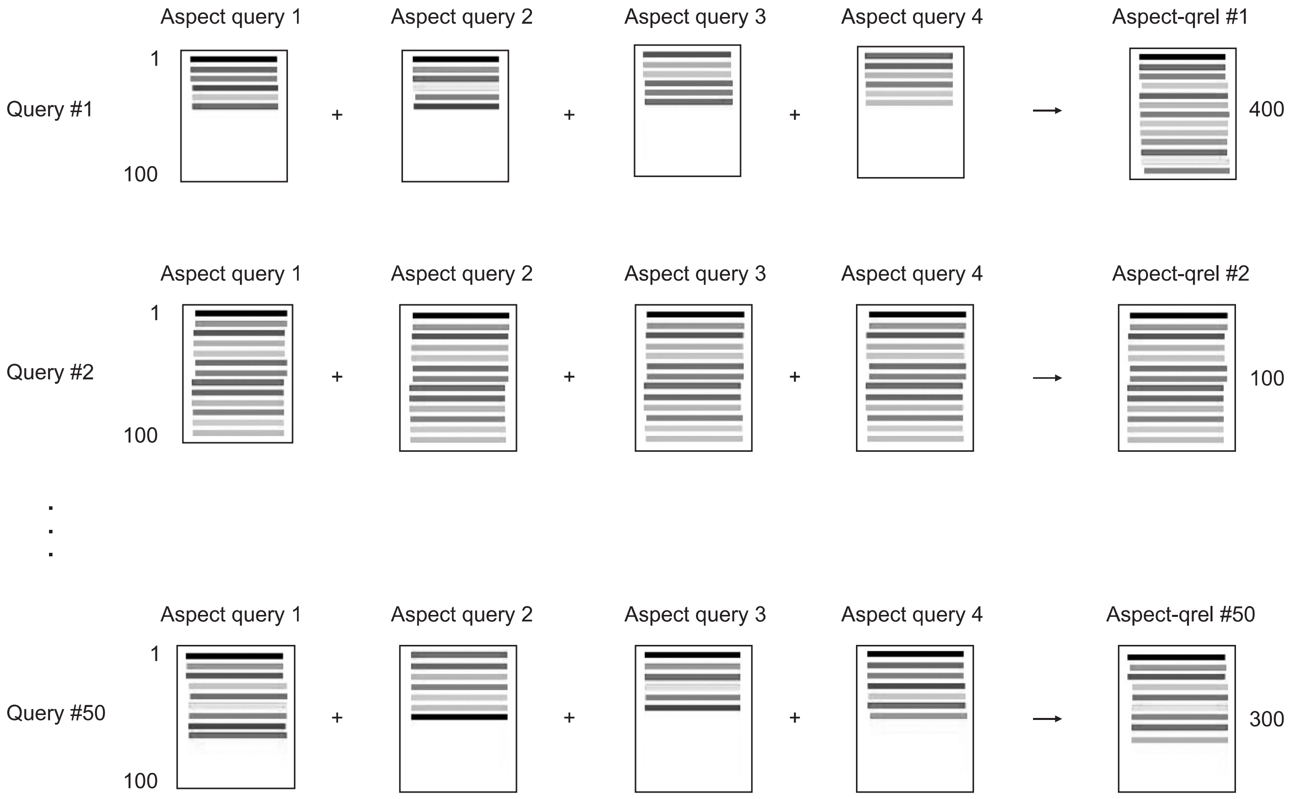

Figure 2. Methodological diagram for generating aspect-qrels.

performed under p-value < 0.01.

III. Results

In this section, we primarily report the experimental results for performance computed by different document-query relevance qrels. Submitted run results from 46 teams participated in 2004 genomics track were supported by NIST, the organization of TREC conference. In our approach, the rank ordering computed by MAP was used as baseline. This baseline was used as the reference for comparing evaluation performance.

1. 2004 Genomics Track Runs

The goal of the TREC genomics track is to create test collections for evaluation of information retrieval and related tasks in the genomics domain. In 2004 genomics track, several teams submitted their retrieval runs and the overall results were released in online after the competition. Ad hoc task, which data was used in our experiment was a standard ad hoc retrieval task using topics obtained from real biomedical research scientists and document from a large subset of the MEDLINE bibliographic database. The document collection for ad hoc retrieval task was a 10-year subset of MEDLINE. Based on MAP, ranking order of participated teams was presented. Runs were calculated using the trec_eval program, a standard scoring system for TREC. To figure out whether aMAP computed by judgment-free aspect-qrels correlates with MAP computed by original human-judged TREC qrels, we formulated rank orderings for runs according to priority of aMAP in descending order. Given a set of 46 IR runs, we rank systems from best-to-worst-performing with respect to MAP and aMAP made through this study. Figure 3 shows system rankings based on aMAPs using aspect query qrels and MAP using TREC qrel for each runs.

2. Rank Correlation

In our experimental result, Spearman's 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 values of whole correlations were 0.6694 in tau and 0.831 in rho under p-value < 0.01. Figure 4 shows tau

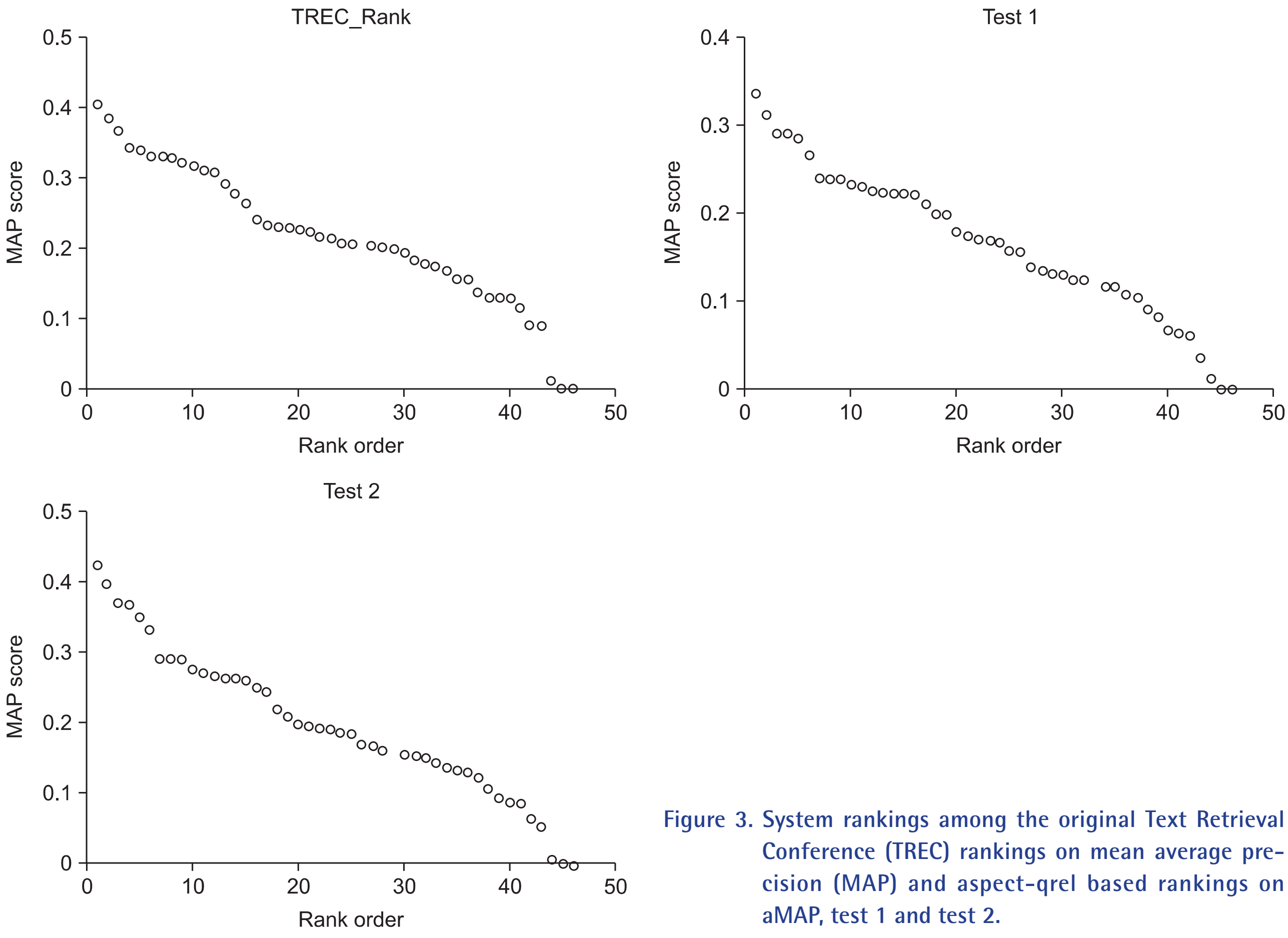

Figure 3. System rankings among the original Text Retrieval Conference (TREC) rankings on mean average precision (MAP) and aspect-qrel based rankings on aMAP, test 1 and test 2.

Table 2. Rank correlations with MAP computed by TREC qrel and aMAP computed by aspect query based 15 qrels

Aspect query based qrels	Kendall's tau	Spearman's rho
A	0.632	0.791
B	0.647	0.813
C	0.71[a]	0.863[a]
D	0.666	0.827
E	0.648	0.818
F	0.65	0.816
G	0.67	0.833
H	0.642	0.81
I	0.674	0.837
J	0.707	0.858
K	0.708	0.862
L	0.662	0.825
M	0.651	0.821
N	0.71	0.861
O	0.664	0.83

Each correlation was calculated under $p < 0.01$.

MAP: mean average precision, aMAP: aspect-qrels MAP, TREC: text retrieval conference.

[a]The highest correlation value.

and rho rank correlations between system rankings obtained by the official TREC MAP and aspect query based on aMAP.

IV. Discussion

In our study for appropriate method to generate a gold standard set in the biomedical retrieval, we supposed to look into the core effects of aspect query method used for constructing relevance judgment without human assessments.

1. The Effect of Aspect Pool Depth on Correlation

Aspect queries suggested to be a source for making relevance judgment are used by taking the top N documents from retrieved pool. In this process, the tunable parameter N was considered to be 100 for the first time. Following TREC terminology we called the number of N to be an aspect pool depth and analyzed the number of pool depth for its correlation. We hypothesized the correlation would be increased as the pool depth becomes deeper.

Figure 5 illustrates that most of data correlations were increased highly within the range of 50 to 100 documents. According to this result, we can argue that the tunable parameter should be in the range of 50 to 100 for composing pseudo-relevant documents.

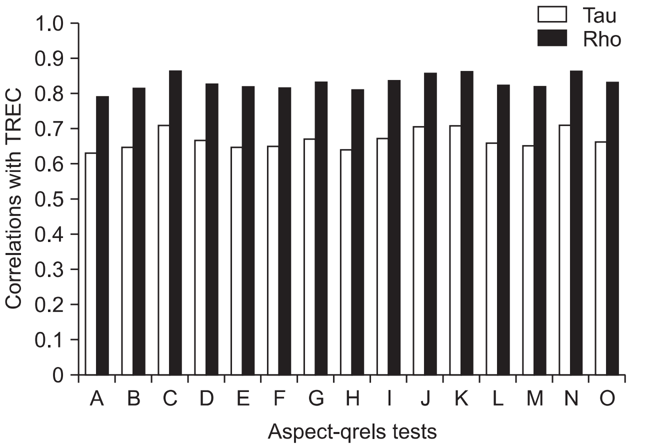

Figure 4. Experimental results for rank correlations among aspect-qrels: Kendall's tau and Spearman's rho coefficients. TREC: Text Retrieval Conference.

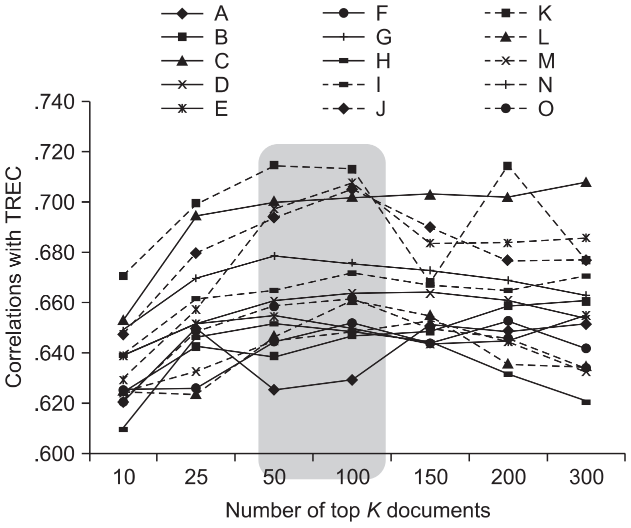

Figure 5. The effect of the number of documents collected per aspect during aspect-qrel creation. TREC: Text Retrieval Conference.

2. The Effect of the Number of Aspect Queries on Correlation

The number of aspect queries to get 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 is any regular formula within the number of queries. According to this result, we can argue that there should be needed to figure the relationship out for a definite number of aspect queries.

3. Comparative Analysis between Single and Union of Aspect Queries

We contextualized the relationship among the singles queries, the average correlation with four aspect queries and the union of four aspect queries. The average for single queries was calculated an arithmetic mean. Figure 6 illustrates that the rho correlations on the union of four aspect queries showed higher correlations than arithmetic average value in principle. Clearly, correlations computed by aspect-qrels perform higher than others (Table 3).

4. Baseline Correlation among Different Retrieval Models

In our approach, BM25 retrieval model was applied as our "seed system" to retrieve queries and document, also produce aspect-qrel files. Our seed system used Porter's stemmer and stop word removal. Each runs generated on BM25 for the analysis of aspect query method.

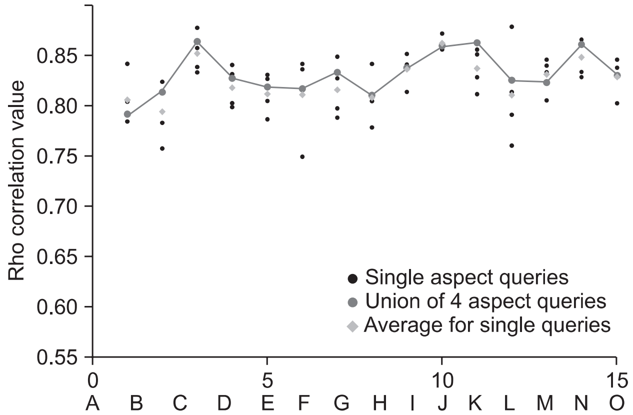

Figure 6. Rho correlation between MAP and aMAP calculated from individual aspect qrels. MAP: mean average precision, aMAP: aspect-qrels MAP.

One might oppose whether the correlations we have observed in this study are high simply because the BM25 seed system performs quite well. We tested on three retrieval models including BM25 probabilistic model, Hiemstra's language model and PL2 based on Poisson estimation for randomness theory. With this experiment, a more pronounced influence appeared and the bias that BM25 model is superior on the performance so as to make the results on aspect query better than assumption turned out to be incorrect (Table 4).

5. Future Implications on Medical Informatics Point of View

In the medical domain, the most frequent use of information searching is to retrieve bibliographic information. As

Table 4. Baseline correlations used different seed systems: BM25, Language model, and PL2

	BM25	Hiemstra's language model	PL2
Kendall's tau	0.678	0.688[a]	0.645
Spearman's rho	0.849	0.86[a]	0.821

Each correlation was calculated under $p < 0.01$.

[a]The highest correlation value.

Table 3. Rho correlation between MAP and aMAP calculated from individual aspect qrels

Aspect-qrel	A1	A2	A3	A4	Union (4)	Average
A	0.841[a]	0.792	0.803	0.784	0.791	0.805
B	0.814[a]	0.782	0.823	0.757	0.813	0.794
C	0.877[a]	0.857	0.833	0.838	0.863	0.85125
D	0.840[a]	0.798	0.802	0.831	0.827	0.81775
E	0.826	0.804	0.786	0.830[a]	0.818	0.8115
F	0.816	0.841[a]	0.836	0.749	0.816	0.8105
G	0.848[a]	0.827	0.788	0.797	0.833	0.815
H	0.841[a]	0.804	0.808	0.778	0.810	0.80775
I	0.839	0.851[a]	0.841	0.813	0.837	0.836
J	0.871[a]	0.860	0.860	0.855	0.858	0.8615
K	0.811	0.851	0.828	0.855	0.862[a]	0.83625
L	0.878[a]	0.760	0.813	0.791	0.825	0.8105
M	0.845[a]	0.839	0.805	0.832	0.823	0.83025
N	0.864	0.865[a]	0.833	0.828	0.861	0.8475
O	0.845[a]	0.837	0.802	0.828	0.830	0.828

Each correlation was calculated under $p < 0.01$.

MAP: mean average precision, aMAP: aspect-qrels MAP.

[a]The highest correlation value of each line.

well as the specificity of MEDLINE documents, it is needed to develop more sophisticated way of medical information retrieval. Test collection is used to evaluate the search system performance and field experts can generate the document-query gold standard. We grounded our hypothesis that it could be possible with using multiple aspect query method on creating biomedical test collections without undue labor intensive task. We examined with the experiment on ranking systems by aspect query based gold standards generated without human experts and its correlations between rankings by assessor-judged gold standards. Finally, we expect that our verifying experiment on using multiple queries to build biomedical test collection could promote medical information retrieval research.

Aspect query, as the elaboration, specification, or paraphrase of the original topic, was generated manually and focused to build document-query relevance by collection top-ranked documents from the retrieved pool. According to our experimental results, 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 y reason of rank ordering of IR systems correlates highly with rankings based on human relevance judgments. It is expected to contribute medical information retrieval, evaluation study in particular.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

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).

References

1. Sanderson M, Braschler M. Best practices for test collection creation and information retrieval system evaluation. Pisa, Italy: TrebleCLEF; 2009. Technical report no.: D4.2.
2. Voorhees EM, Tice DM. Building a question answering test collection. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000. p.200-7.
3. Oard DW, Soergel D, Doermann D, Huang X, Murray GC, Wang J, Ramabhadran B, Franz M, Gustman S, Mayfield J, Kharevych L, Strassel S. Building an information retrieval test collection for spontaneous conversational speech. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004. p.41-8.
4. Hersh W, Buckley C, Leone TJ, Hickam D. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994. p.192-201.
5. Heppin KF. MedEval: a Swedish medical test collection with doctors and patients user groups. In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Document, 2010. p.1-7.
6. Efron M. Using multiple query aspects to build test collections without human relevance judgments. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, 2009. p.276-87.
7. Sanderson M, Joho, H. Forming test collections with no system pooling. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004. p.33-40.
8. Soboroff, I, Nicholas, C, Cahan, P. Ranking retrieval systems without relevance judgments. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001. p.66-73.
9. Wu S, Crestani F. Methods for ranking information retrieval systems without relevance judgments. In: Proceedings of the 2003 ACM Symposium on Applied Computing, 2003. p.811-6.
10. Grady C, Lease M. Crowdsourcing document relevance assessment with mechanical turk. In: Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010. p.172-9.
11. Cao YG, Ely J, Antieau L, Yu H. Evaluation of the clinical question answering presentation. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, 2009. p.171-8.
12. Luo G. Design and evaluation of the iMed intelligent medical search engine. In: Proceedings of the IEEE International Conference on Data Engineering, 2009. p.1379-90.

13. National Institute of Sandards and Technology (NIST). Text retrieval conference (TREC) [Internet]. Gaithersburg (MD): NIST; c2012 [cited at 2011 Oct 17]. Available from: http://trec.nist.gov/.
14. Si L, Lu J, Callan J. Combining multiple resources, evidence and criteria for genomic information retrieval. In: Proceedings of the Fifteenth Text Retrieval Conference (TREC), 2006.
15. Yin X, Huang X, Li Z. Promoting ranking diversity for biomedical information retrieval using wikipedia. In: Proceedings of the 32nd European Conference on Advances in Information Retrieval, 2010. p.495-507.
16. Yin X, Huang JX, Zhou X, Li Z. A survival modeling approach to biomedical search result diversification using wikipedia. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010. p.901-2.
17. National Science Foundation Information Technology Research. TREC genomics track [Internet]. Arlington (VA): National Science Foundation Information Technology Research; c2008 [cited at 2011 Oct 17]. Available from: http://ir.ohsu.edu/genomics/.
18. Hersh WR. Report on the TREC 2004 genomics track. ACM SIGIR Forum 2005;39:21-4.
19. Korean Medical Library Engine [Internet]. Seoul, Korea: Korean Medical Library Engine; c2011 [cited at 2011 Jul 20]. Available from: http://www.kmle.co.kr/.
20. National Library of Medicine. Medical Subject Headings [Internet]. Bethesda (MD): National Library Medicine; c2011 [cited at 2011 Jul 20]. Available from: http://www.nlm.nih.gov/mesh/MBrowser.html.