Document Re-ranking by Generality in Bio-medical Information Retrieval

Xin Yan[1], Xue Li[1], and Dawei Song[2]

[1] School of Information Technology and Electrical Engineering, University of Queensland, ITEE, University of Queensland, QLD 4072, Australia
{yanxin, xueli}@itee.uq.edu.au
[2] Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
dawei_song2005@hotmail.com

Abstract. Document ranking is an important process in information retrieval (IR). It presents retrieved documents in an order of their estimated degrees of relevance to query. Traditional document ranking methods are mostly based on the similarity computations between documents and query. In this paper we argue that the similarity-based document ranking is insufficient in some cases. There are two reasons. Firstly it is about the increased information variety. There are far too many different types documents available now for user to search. The second is about the users variety. In many cases user may want to retrieve documents that are not only similar but also general or broad regarding a certain topic. This is particularly the case in some domains such as bio-medical IR. In this paper we propose a novel approach to re-rank the retrieved documents by incorporating the similarity with their generality. By an ontology-based analysis on the semantic cohesion of text, document generality can be quantified. The retrieved documents are then re-ranked by their combined scores of similarity and the closeness of documents' generality to the query's. Our experiments have shown an encouraging performance on a large bio-medical document collection, OHSUMED, containing 348,566 medical journal references and 101 test queries.

Keywords: Generality, Relevance, Document Ranking.

1 Introduction

Document ranking is a fundamental feature for information retrieval (IR) systems. In general, an IR system ranks documents based on how close or relevant a document is to a query. In traditional models such as vector space model, documents are represented by vectors of keywords. The relevance is computed based on similarity (often defined by functions such as cosine or inner product) between the document and the query vectors.

Due to information explosion and popularity of WWW information retrieval, however, the sufficiency of using relevance alone to rank documents has been questioned by the generality retrieval problem.

M. Kitsuregawa et al. (Eds.): WISE 2005, LNCS 3806, pp. 376–389, 2005.
© Springer-Verlag Berlin Heidelberg 2005

On one hand, information explosion somehow increases not only the quantity of information but also the variability. For instance, consider a topic for general AIDS information in PubMed[1], a medical searching service. Thousands of documents may be retrieved in a wide range such as treatment, drug therapy, transmission, diagnosis and history. User may need to have a holistic view on the topic by first reading some general and conclusive documents to find something reasonably related to their information needs. This is a challenge to the traditional relevance based document ranking since it cannot help user to sort out the relevant documents which are also general in content.

Moreover, the growing popularity of WWW information retrieval makes domain-specific information retrieval open to the public. For example there are human identified and labeled documents about patient education available in PubMed. Other patient education materials are separately maintained on WWW such as MedicineNet[2]. Easy-to-understand and jargon free information is needed by user with little domain knowledge. However, current ranking mechanism does not focus on this perspective.

Based on the concerns of generality retrieval, we argue that the factor "generality" should be taken into account in document ranking process. We need to consider both document and query generality, which separately refers to how general it is for a document/query to describe a certain topic. The goal of this research is to improve the query performance of domain specific (bio-medical literature in this paper) information retrieval by re-ranking retrieved documents on generality.

A novel ontology based approach to the calculation of generality is developed via analyzing the semantic cohesion of a document. The documents are then ranked by a combined score of relevance and the closeness of documents' generality to the query's. Experiments have been conducted on a large scale bio-medical text corpus, OHSUMED, which is a subset of MEDLINE collection containing 348,566 medical journal references and 101 test queries. Our approach has demonstrated an encouraging improvement on performance.

The remainder of this paper is organized as follows: Section 2 presents related work. Our methods to re-rank documents on generality are proposed in Section 3. Section 4 reports the experimental results. Section 5 finally concludes the paper and addresses future research directions.

2 Related Work

To the best of our knowledge, the previous studies on generality in IR literature [1], [2], [3], [4], [5] focus on two different aspects: query generality (i.e. query scope) and content-based document generality. Since our proposed re-ranking method is a process comparing the closeness of document's generality scores to the queries, it is necessary to review both literatures.

[1] http://pubmed.gov
[2] http://www.medicinenet.com

2.1 Query Generality

The studies [2], [3], [4] about query generality mainly focus on the overall generality of retrieval rather than the generality of individual documents. Van Rijsbergen [4], [3] regarded query generality as "a measure of the density of relevant documents in the collection". Derived from Van Rijsbergen's definition, Ben He [2] defined query generality as follows:

$$\omega = -log(\frac{N_Q}{N}) \tag{1}$$

where N_Q is the total number of documents containing at least one query term and N is the total number of documents in the collection.

However, because of the content variability of retrievals, it is not sufficient to quantify the query generality purely based on the statistical methods. Let's consider two topics T_1 and T_2. T_1 requires literature reviews about AIDS, T_2 requires reviews about SARS, a newly discovered disease. In PubMed, a boolean query to get all the review articles about "AIDS" may result in 7650 documents. Whereas, there are only 220 review articles about "SARS". Moreover, the term "SARS" appears in 3394 references but "AIDS" appears in 108439. Is the query for T_1 more general than T_2? The answer is probably "no", because "SARS" is a new disease which has less related documents in collection than "AIDS".

2.2 Document Generality

The studies about document generality aim at finding approaches to rank general documents more closely to the query. Allen [1] argued that user has needs to know whether a document retrieved is general or concrete. Document generality was defined as the mean generality of terms in the documents. The generality of 64 words was determined. Those words were used to form a reference collection. Half of the words in the collection were regarded as general and the other half as concrete. Joint entropy measure was used to verify that general terms were more related to each other than concrete terms. Thus. through the relatedness computation between the terms in documents and those 64 terms in the reference collection, the generality of terms in the documents could be calculated.

However, some problems still remain unsolved. First, in [1], the generality of documents was judged manually for the purpose of evaluation. However, for dealing with large collections, this is obviously not practical. Secondly, not only the statistical term relatedness, but also the semantic relationships between terms need to be taken into account. Sometimes general terms may have low relatedness if they are not in a same domain. In the area of bio-medical information retrieval, for example, a stomach medicine may be semantically related to a skin medicine in terms of their generality. However, they may not have a statistical relatedness simply due to no co-occurrence in the text corpus. Thirdly, how to apply generality ranking to improve IR performance has not been discussed. Moreover, we need to consider how to combine the relevance ranking with generality ranking. User does not want a document with high generality but very low relevance to query.

The study of subtopic retrieval [5] addressed that there is a need (e.g literature survey) to find documents that "cover as many different subtopics of a general topic as possible" [5]. The subtopic retrieval method solved two problems we referred above. Firstly, a new evaluation framework was developed to evaluate the performance of re-ranking. The subtopic recall and precision can be calculated for every retrieved document since the human assigned subtopic labels are available for those documents in TREC interactive track. Documents with high generality will have a good balance between subtopic recall and subtopic precision. This framework avoids the human judgement of document generality. Secondly, the relevance ranking has been considered when re-ranking documents by generality.

There are still some major differences between the study of subtopic retrieval [5] and our proposed approach.

1. We assume that the relevance judgment of a document in OHSUMED collection is independent to that of the others. In the study of subtopic retrieval, relevance between two documents may depend on which document user will see in the first time.
2. Semantics inherent in the documents is considered in our research. We measure the ontology based semantic relationships of concepts in document in order to compute generality. In the study of subtopic retrieval, only statistical methods were used.

3 Proposed Approach

In this section, we first define our research problem and give the intuition of our solution, followed by the detailed computational methods.

The research problem is defined as: given a ranked list of documents R retrieved by a query Q, find R' so that documents in R' are ordered by both their relevance to Q and the closeness of their generality to Q's.

We approach the generality ranking problem from two perspectives. The first is to consider the query generality. We believe that generality ranking depends on both query generality and document generality. To a specific query (i.e., a query with low generality), it is not proper to simply rank general documents higher than the specific ones. The second is to consider the semantics in documents. For instance, a stomach medicine is not statistically but semantically same as a skin medicine in terms of their generality.

A query can be regarded as a short document. So in the same way a query is to be computed for its generality as if a document. Then the documents are re-ranked by comparing the closeness of documents' generality scores to the query's.

On the other hand, the semantics of documents can be computationally gripped in terms of ontology. In our work, we use bio-medical documents together with an ontology database called MeSH hierarchical structure (or MeSH tree) in bio-medical domain. Our purpose is to compute generality of text by considering the semantic properties and relations of terms appearing in the MeSH

tree. For example, stomach medicine and skin medicine both belong to "Chemicals and Drugs" no matter how different their usages are. Here we regard the terms in text which can be found in MeSH ontology as domain specific concepts or MeSH concepts. The terms in text which cannot be found in MeSH ontology are referred to non-ontology concepts.

We introduce cohesion as a key feature of generality. When there is a focused topic or theme discussed in a document, the terms are closely correlated in a certain context. The cohesion of a document is regarded as a computation of the associations between the concepts found in the MeSH tree. It reflects the frequencies of the associated concepts that appear in the MeSH ontology. The more closely the concepts are associated, the more specific the document is.

In following subsections, we will describe the MeSH hierarchical structure and propose a method to identify MeSH concepts from text. We then present our approach to computational generality of documents.

3.1 MeSH Hierarchical Structure

All the headings used to index OHSUMED documents are well organized in a hierarchical structure namely MeSH tree. Figure 1 is a fragment of the MeSH tree.

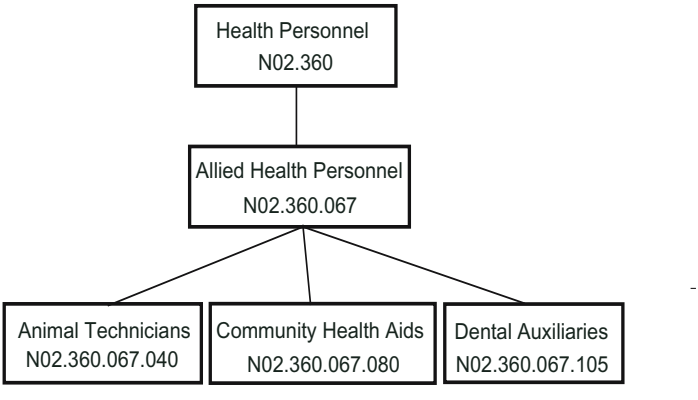


Fig. 1. A Fragment of MeSH tree

The MeSH terms are numbered and organized based on a broader/ narrower relationships in the tree. In this example, the heading "Allied Health Personnel" is a kind of "Health Personnel" and "Community Health Aides" is a kind of occupation under "Health Personnel".

Moreover, MeSH provides entry terms which may act as synonyms of a certain heading. In the given document example, the heading "Allied Health Personnel" has following entry terms: "Allied Health Personnel", "Allied Health Paramedics", "Paramedical Personnel", "Specialists, Population Program" and

"Paramedics". With entry terms, it is possible to take advantage of semantic relation between terms to identify synonyms.

3.2 Concept Identification Algorithm

In order to use MeSH ontology to extract the semantic relations between terms, the MeSH concepts in the text corpus must be recognized. The proposed algorithm of concept identification aims to allocate a single word or a compound(noun) from the corpus as a concept in the MeSH tree.

The major problem that the algorithm is concerned about is: a part of a compound term may match with a MeSH concept. For example, the compound "Plant Viruses" contains the term "Viruses". If we stop the concept identification process after a match of "Viruses" in the MeSH tree is found, then "Plant" will be mistakenly regarded as a term not in domain ontology. Indeed, "Plant Viruses" is also a MeSH concept. We solve the problem by introducing the conceptual marking tree (CMT) that is derived from the MeSH tree. The structure of a node in CMT is shown in Figure 2. A concept C is a sequence of terms $\{T_1 \dots T_n\}$, where n is the length of C. The occurrence information of individual terms are stored separately in the cells of an array. In cell T_i, $0 \leq i \leq n$, we use P_i to store a set of position values $\{p_{i1} \dots p_{im}\}$, where m is the term frequency of T_i in a document. p_{ij} $(0 \leq j \leq m)$ is the term position of the jth occurrence of T_i. The term position p_{ij} indicates that there are $(p_{ij} - 1)$ terms before T_1 from the beginning of a document.

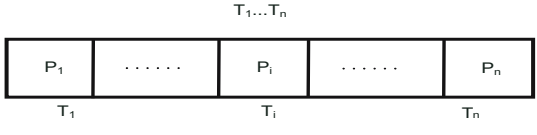


Fig. 2. Data Structure of a Node in CMT

There are 3 steps to perform the conceptual marking for a document.

1. Pick up a term t which is the k-th term counted from the beginning of the document (initially $k = 0$).
2. Locate t in CMT.
3. Assign the position value k to p_{ij} in P_i. j will be increased by one automatically when a new element is added to P_i.
4. Increase k by one, then go to step 1.

For example, the following is a one-sentence document just containing one sentence:

```
Over 390 individual descriptions of plant viruses or virus
groups are provided.
```
[3]

[3] http://www.dpvweb.net/dpv/index.php

In this example, "plant viruses" and "viruses" are all MeSH concepts. We assume that stemming has been done so that "viruses" can be identified as "virus". After the CMT is created for this document, the concept "plant viruses" in CMT have two cells, T_1 = "plant", T_2 = "viruses". $p_{11} = 6$, $p_{21} = 7$, $p_{22} = 9$. The concept "viruses" has one cell T_1 = "viruses" where $p_{11} = 7$, $p_{12} = 9$.

After marking CMT, if it is always true that $p_{(i-1)j} = p_{(i)j} + 1$ $(1 \leq i \leq m)$, then the concept C is identified as a candidate concept at its jth occurrence in the document. If in the same place of the document, no other candidate concepts can be found with more compound terms than concept C, then C is identified as the concept at its jth occurrence in the document. For the above example, we may find that the MeSH concept "viruses" may be identified as the candidate concept in position 7 and 9. However, the concept "plant viruses" has $p_{11} = p_{21} + 1$. Furthermore, it has two constituent terms but the concept "viruses" only has one. Thus it is "plant viruses" rather than "virus" which is identified as the concept at position 6.

3.3 Document Cohesion

With MeSH hierarchical structure (tree), it is possible to retrieve the semantic distance between MeSH concepts according to their positions in tree.

We introduce the concept of document cohesion which is a state or quality that the elements of a text (e.g. clauses) "tend to hang together" [6]. The intuition of our approach is based on a hypothesis that document with less cohesion would be more general. Consider the following two definitions of Severe Acute Respiratory Syndrome (SARS). Definition 1 comes from an FAQ page of Centers for Disease Control and Prevention (CDC) [4] under a section namely "what everyone should know". Definition 2 is an official definition from the Department of Health in Hong Kong [5]. Obviously, definition 1 is more general than definition 2.

1. Severe Acute Respiratory Syndrome(SARS) is a viral respiratory illness that was recognized as a global threat in March 2003, after first appearing in Southern China in November 2002.
2. Severe Acute Respiratory Syndrome(SARS) is a viral respiratory infection caused by a coronavirus (SARS-CoV).

In definition 2, four MeSH concepts can be identified: "Severe Acute Respiratory Syndrome", "respiratory infection", "coronavirus" and "SARS-CoV". In definition 1, "Severe Acute Respiratory Syndrome" and "China" are identified as MeSH concepts.

What makes definition 1 be more general than definition 2? We found that there is stronger cohesion in definition 2 than in definition 1. In other words, concepts in definition 2 are more strongly associated than those in definition 1. "Severe Acute Respiratory Syndrome" is a kind of "respiratory infection" in terms of MeSH ontology. Moreover, "SARS-CoV" is a kind of "coronavirus".

[4] http://www.cdc.gov/ncidod/sars/faq.htm
[5] http://www.info.gov.hk

However, in definition 1, there is not a direct relationship between 'Severe Acute Respiratory Syndrome" and "China" in terms of the MeSH ontology.

Following the above observations, it seems that the document generality is somehow related to document cohesion. The higher a document's degree of cohesion, the lower its generality. In our research, the degree of document cohesion is inversely proportional to the mean semantic distance of all the pairs of concepts in document. The calculation of semantic distance is based on the Leacock-Chodorow similarity [7] function which measures the shortest path between two concepts in the MeSH tree. We adopt Leacock-Chodorow similarity and propose our algorithms to compute cohesion.

$$Cohesion(d_i) = \frac{\sum_{i,j=1}^{n} Sim(c_i, c_j)}{Numberof Associations}, \ where \ n > 1, \ i < j \tag{2}$$

$$Sim(c_i, c_j) = -log\frac{len(c_i, c_j)}{2D} \tag{3}$$

$$Numberof Associations = \frac{n(n-1)}{2} \tag{4}$$

In Equation 2, n is the total number of MeSH concepts in a document d_i. $Sim(c_i, c_j)$ is a function computing the Leacock-Chodorow semantic similarity by using the shortest path $len(c_i, c_j)$ between c_i and c_j in the MeSH tree. $Numberof Associations$ is the total number of associations among different MeSH concepts, which is defined in Equation 4.

In Equation 3, D is the maximum MeSH tree depth. In our experiments, D is 11. The scope of Equation 2 is $[0, -log(\frac{1}{22})]$. As to a document without any MeSH concepts or with only one MeSH concept, its document cohesion is 0. For a documents with strongest associations among all the concepts within the document, its cohesion is $-log(\frac{1}{22})$, the maximum value.

3.4 Document Generality

We give Equation 5 for calculating document generality. In Equation 5, $DG(d_i)$ denotes the generality of a document d_i.

$$DG(d_i) = \frac{1}{Cohesion(d_i) + 1} \tag{5}$$

The query generality computation is similar to the computation of document generality. The difference between them is that we take Equation 1, with a new name Statistical Query Generality (SQG) as an option for query generality calculation.

$$QG = \frac{SQG}{Cohesion(Q) + 1} \tag{6}$$

In Equation 6, QG is the query generality. The calculations of query cohesion is the same as document cohesion.

However, we believe that it is better to give high ranks to those documents whose generality are close to the queries'. For example, it is not suitable to give high ranks to the review or introduction papers on "malignant pericardial effusion" for the query "best treatment of malignant pericardial effusion in esophageal cancer". Thus, we rank the documents by comparing the closeness of documents' generality scores to the query's. In this research the generality closeness between query Q and document d_i is computed as the absolute value of the difference between $DG(d_i)$ and QG.

3.5 Combining Relevance and Generality

As an important step in our proposed approach, we consider both the document relevance and generality. Here we treat information retrieval system as a black box. Through the query submitted as input, the output of the black box is a ranked list where documents are scored. Let $RScore(d_i)$ denote the relevance score given to a ranked document d_i and QG is the query generality. The final score considering both document relevance and generality is given in the following formula.

$$Score(d_i, Q) = RScore(d_i)^{\alpha} * e^{-|DG(d_i) - QG|^{\beta}} \tag{7}$$

α and β are parameters for a well tuned performance.

4 Experiments and Evaluations

4.1 Data Set

Our experiments are designed based on the OHSUMED corpus, which is a subset of Medline and contains 348566 medical references. There are a number of fields in a reference, such as title, abstract, author, source and publication type. In our research, we use title and abstract only. The following is a fragment of a sample document in OHSUMED collection, where

- .I Sequential Identifier
- .U MEDLINE identifier (UI)
- .T Title (TI)
- .P Publication type (PT)
- .W Abstract (AB)
- .A Author (AU)
- .S Source (SO)

```
.I 1
.U 87049087
.S Am J Emerg Med 8703; 4(6):491-5
.T Refibrillation managed by EMT-Ds: incidence and
outcome without paramedic back-up.
.P JOURNAL ARTICLE.
.W Some patients converted from ventricular
fibrillation to organized rhythms by
defibrillation-trained ambulance technicians (EMT-Ds)
will refibrillate before hospital arrival.
......
.A Stults KR; Brown DD.
```

4.2 Queries

In OHSUMED there are 106 topics and their relevance judgments made by novice physicians. Each topic has two parts: the patient information and the physician's information need. In this research, 106 test queries are formed by combining both parts for each of the 106 topics. In addition, queries 8, 28, 49, 86, and 93 are dropped for there are no relevant documents identified for them. Therefore, a total number of 101 test queries are used in our experiments.

There are queries apparently asking for review information. The following eight review-type queries are selected to test the effect of query generality.

- No.4 reviews on subdurals in elderly
- No.11 review article on cholesterol emboli
- No.17 RH isoimmunization, review topics
- No.31 chronic pain management, review article, use of tricyclic antidepressants
- No.34 review article on adult respiratory syndrome
- No.54 angiotensin converting enzyme inhibitors, review article
- No.105 review of anemia of chronic illness
- No.106 HIV and the GI tract, recent reviews

4.3 Baseline and Pre-processing

In our experiments, Lucene[6] is used as the baseline IR system to index and retrieve the titles and abstracts of documents in OHSUMED collection. All terms are filtered by the SMART 571 stop word list and stemmed using the Porter stemming algorithm. The MeSH concepts are identified by using our conceptual marking tree algorithm.

[6] http://lucene.apache.org/java/docs/index.html

4.4 Evaluation Methodology

In our experiments, the baseline system is used to retrieve 1000 documents for each test query. We then cover all three possible cases where query generality, document generality and SQG are used solely or together in a reasonable manner. Those three cases are derived from our proposed Equation 5, 6 and 7 for re-ranking the documents retrieved by the baseline.

Case One: DG. The first case is to re-rank documents without considering query generality. The scoring function (Equation 7) is changed to Equation 8:

$$Score(d_i, Q) = RScore(d_i)^{\alpha} * DG(d_i)^{\beta} \tag{8}$$

α and β are parameters for a well tuned performance.

Case Two: DG+QG+SQG. The second case is to re-rank documents by considering both document generality and query generality with SQG. This option exactly consists of Equation 5, 6 and 7.

Case Three: DG+QG-SQG. The third case is to re-rank documents by considering both document generality and query generality without SQG. The Equation 6 is changed to Equation 9 to remove SQG. Equation 9 is then used for calculating QG in Equation 7.

$$QG = \frac{1}{Cohesion(Q) + 1} \tag{9}$$

4.5 Performance Indicators

The performance of re-ranking are measured in two ways. Firstly we compare the precision and recall of re-ranking with the original ranking given by baseline system for all the 101 test queries. Secondly, we check if all the review type queries get larger improvement in term of average precision.

4.6 Evaluations

Table 1 gives detailed precisions of each algorithm at different recall levels averaged over 101 test queries. In Table 2, we show the performance of the algorithms on the review type queries. The mean average precision ("MAP" in the tables) and the percentages of improvement in MAP ("%" in the tables) are summarized.

It seems a general case that DG+QG+SQG and DG+QG-SQG improve the query performance for the 101 queries over the baseline. However, DG degrades overall query performance slightly. Therefore, it is more effective to re-rank documents based on the closeness between document and query generality rather than considering document generality alone.

Table 1. Detailed Precision-Recall Comparisons

R	Baseline	DG	DG+QG+SQG	DG+QG-SQG
0	0.6369	0.6311	**0.6512**	0.6366
0.1	0.4071	**0.4110**	**0.4139**	**0.4089**
0.2	0.3239	0.3222	**0.3281**	**0.3263**
0.3	0.254	0.2480	**0.2635**	**0.2644**
0.4	0.1963	0.1942	**0.1985**	**0.2024**
0.5	0.1679	**0.1681**	**0.1689**	0.1675
0.6	0.1396	0.1364	0.1389	0.1372
0.7	0.088	0.0873	**0.0884**	**0.0893**
0.8	0.0544	0.0537	0.0543	0.0542
0.9	0.0223	0.0221	**0.0224**	0.0218
1	0.0018	0.0018	**0.0019**	0.0017
MAP	0.1849	0.1834	**0.1883**	**0.1873**
%		-0.81%	**1.84%**	**1.30%**

Table 2. Precision Improvement on Review Type Queries

QNo.	Baseline	DG	DG+QG+SQG	DG+QG-SQG
4	0.0821	0.0450	**0.0824**	0.0375
11	0.0741	0.0712	**0.1046**	**0.0936**
17	0.0021	0.0020	0.0021	**0.0023**
31	0.1522	0.1415	**0.1542**	**0.1544**
34	0.0193	**0.0198**	0.0188	0.0164
54	0.1099	**0.1125**	0.1025	**0.1181**
105	0.2950	**0.3003**	0.2926	0.2789
106	0.0085	0.0053	**0.0121**	0.0066
MAP	0.0929	0.0872	**0.0962**	0.0885
%		-6.14%	**3.55%**	-4.76%

Unlike documents, queries are normally very short. Consequently there is less information involved in the computation of query cohesion, which in turn may not be sufficient enough to reflect query generality. SQG is therefore an important complementary component for effectively measuring query generality. This is demonstrated by the fact that the DG+QG+SQG, which takes SQG into account, always performs better than DG+QG-SQG.

In summary, as both document generality and query generality with SQG are considered, DG+QG+SQG performed the best to benefit generality retrieval.

Moreover, Table 2 shows the better performance of the DG+QG+SQG algorithm on review type queries. There is an encouraging 3.55% improvement over the baseline. We performed a dependent t-test (Paired Two Sample for Means) which compares the paired precisions between the baseline and the DG+QG+SQG algorithm over different queries in Table 2. With a $p - value$ less than 0.05, it turns out that the improvement is significant. This also verifies our motivation discussed in Section 1 that the generality retrieval happens more often for review type queries from non-domain-expert user.

5 Conclusions and Future Work

In this paper, we studied a generality retrieval problem in bio-medical area where document ranking is based not only on relevance but also on generality. Traditional document ranking methods are insufficient for generality retrieval because they are depends on relevance only. This paper argued that the "generality" is an important complement to the traditional relevance based ranking. The intuition is that when search results are returned by IR system, user, especially non-domain-expert user, may expect to see the general and conclusive documents on the top of the list, so that they can first have an overview on the topic rather than going into the specific technical details directly.

We have proposed a novel ontology-based approach in biomedical IR to re-rank the retrieved documents via generality. Our approach is distinct as to make use of the MeSH ontology structure in bio-medical domain in order to compute the generality from statistical as well as semantic perspectives. Moreover, query generality and document generality were both considered in our proposed re-ranking algorithms. Documents are scored and re-ranked by a combination of their relevance to query and the closeness of documents' generality to the query's. Experiments have been conducted on a large corpus namely OHSUMED. Our approach shows an improved query performance and encourages us to pursue the further investigation. Our approach can also be easily generalized into other domains provided that the domain specific ontologies are available.

We plan to study other factors for ranking document by generality. So far we have considered quantifying only the semantic relationships amongst MeSH concepts in order to calculate the document cohesion. In our further study we will explore other features of document generality and incorporate the relationships between general and domain-specific concepts via statistical approaches, such as term co-occurrence counts.

Acknowledgments

The work reported in this paper has been funded in part by the Co-operative Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Governments CRC Programme (Department of Education, Science and Training).

References

1. Allen, R.B., Wu, Y.: Generality of texts. In: Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology. (2002) 111–116
2. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: 11th Symposium on String Processing and Information Retrieval, Padova, Italy (2004) 43–54

3. Plachouras, V., Cacheda, F., Ounis, I., Rijsbergen, C.v.: University of glasgow at the web track: Dynamic application of hyperlink analysis using the query scope. In: In Proceedings of the 12th Text Retrieval Conference TREC 2003, Gaithersburg (2003)
4. Van Rijsbergen, C.J.: Information Retrieval. London; Boston: Butterworths (1979)
5. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. (2003) 10–17
6. Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics **17** (1991) 21–48
7. Leacock, C., Chodorow, M.: Combining local context and wordnet similarity for word sense identification. In: Fellbaum. (1998) 265C283