Document generality: its computation for ranking

Document Generality: its Computation for Ranking

Xin Yan **Xue Li** **Dawei Song**

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

Abstract

The increased variety of information makes it critical to retrieve documents which are not only relevant but also broad enough to cover as many different aspects of a certain topic as possible. The increased variety of users also makes it critical to retrieve documents that are jargon free and easy-to-understand rather than the specific technical materials. In this paper, we propose a new concept namely document generality computation. Generality of document is of fundamental importance to information retrieval. Document generality is the state or quality of document being general. We compute document generality based on a domain-ontology method that analyzes scope and semantic cohesion of concepts appeared in the text. For test purposes, our proposed approach is then applied to improving the performance of document ranking in bio-medical information retrieval. The retrieved documents are re-ranked by a combined score of similarity and the closeness of documents' generality to that of a query. The experiments have shown that our method can work on a large scale bio-medical text corpus OHSUMED (Hersh, Buckley, Leone & Hickam 1994), which is a subset of MEDLINE collection containing of 348,566 medical journal references and 101 test queries, with an encouraging performance.

Keywords: generality, document ranking, re-ranking

1 Introduction

Generality is the state or quality of being general, according to its definition in Webster Dictionary. A document with high generality might be general or broad in its meaning such as tutorials and reviews. A document with low generality might be specific and narrow in its meaning, for example a journal paper talking about a specific research problem. Generality retrieval is an information searching behavior to find documents which are both relevant to the query and above a certain degree of generality. The trend of generality retrieval is resulted by the information explosion and the popularity of WWW searching. Generality of documents should act as an importance role in information retrieval.

On the one hand, information explosion somehow increases not only the quantity of information but also the variety. For instance a query for general AIDS information in PubMed[1], a medical searching service, may bring some troubles. 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 glance of the topic on the whole, a kind of documents which are not only relevant but also broad enough in meanings to cover as many different aspects of a certain topic as possible, to be retrieved. In this example, user may request review articles of AIDS information.

On the other hand, the growing popularity of WWW information retrieval makes domain-specific information retrieval open to the public. Easy-to-understand and jargon free information is needed by users with insufficient domain knowledge. For example, the patient education materials and tutorials of diseases in bio-medical domain are often requested by the public rather than those materials which are technical and specific.

However, to the best of our knowledge, there is a lack of solutions in literature to satisfy the stringent requirement of generality-based retrieval. The first problem we need to solve is how to compute document generality. In this paper we develop a novel ontology-based document generality computation method via analyzing the scope and semantic cohesion of a document. Our method is then applied to improving the performance of document ranking in bio-medical information retrieval.

Document ranking is well known to be a critical component in information retrieval system. It is the computer judgements of how relevant a document is to a query comparing with other documents retrieved by the same query. Due to the quantity of search result and the limitation of user's time and patience, it is impractical for user to review all the retrieved documents and judge their relevances. In what order to present retrieved documents is a key problem in IR research area.

Based on an assumption that users have a sequential browsing behavior, document ranking determines the presentation order of those retrieved documents. The order is based on how close or relevant a document is to a query. In general, relevance is computed by similarity functions. In traditional IR models such as the vector space model (Salton, Wong & Yang 1975), documents are represented by vectors of keywords and ranked by how similar the document vectors are to the query vector. Two widely used similarity functions are cosine similarity and inner product.

Generality retrieval challenges the traditional document ranking since traditional ranking process is insufficiently based on similarity only. For a simple query "AIDS" in PubMed, we assume that a

[1] http://pubmed.gov

user's information need is to retrieve general information about AIDS. One of the documents retrieved by PubMed, is a specific research paper namely *Multiple Dimensions of HIV Stigma and Psychological Distress Among Asians and Pacific Islanders*. Another article about general AIDS information, *HIV/AIDS: A Minority Health Issue*, is also retrieved. As a result of similarity-based ranking, the former document is ranked much higher than the latter one, whereas, the latter one is closer to user's information need for general AIDS information.

Based on the above discussions, we argue that the factor of "generality" should be taken into account in a document ranking process. Our purpose is to improve the query performance of domain specific (biomedical literature in this paper) information retrieval by re-ranking retrieved documents on generality.

In order to re-rank retrieved documents by generality, we need to know if the generality ranking is required. In practice, there are three ways to determine user's need of generality-based retrieval: manual, semiautomatic and automatic.

Manual Detection of Query Generality User explicitly labels query as general or specific to indicate if the general or specific documents are required.

Semiautomatic Detection of Query Generality User uses a set of pre-defined words such as "review", "introduction" and "tutorial" to test query. User's feedback is needed after retrieval in order to verify user's need of generality retrieval.

Automatic Detection of Query Generality System automatically estimates the generality of query as if it were a document.

In our research, we assume that user's needs of generality retrieval is pre-determined by the IR system through any of the three ways we mentioned above. The focus of our work is to investigate on how to rank documents by their generality.

A novel ontology-based document re-ranking framework is proposed. Based on the hypothesis that there is no dependence between the document generality and its similarity to a query, the documents are ranked by a combined score of similarity and the closeness of documents' generality to the query's. Experiments have been conducted on a large scale biomedical text corpus, OHSUMED (Hersh et al. 1994), which is a subset of MEDLINE collection containing 348,566 medical journal references and 101 test queries. By submitting those queries to our IR baseline system, the similarity of retrieved documents to queries are computed and scored. The correlation analysis between document generality and its similarity score further proves our hypothesis of the independent relationship between generality and similarity. The comparison of retrieval performances before and after re-ranking process reveals that our approach demonstrates an encouraging improvement on technical generality retrieval performance with a positive impact to the overall performance of 101 queries.

The remainder of this paper is organized as follows: Section 2 presents related work. Section 3 gives a detailed definition of generality. Our methods for re-ranking documents on generality are proposed in Section 4. Section 5 reports experimental setup and results. Section 6 concludes the paper and addresses future research directions.

2 Related Work

To our knowledge, no researches directly focusing on generality computation are currently available. The studies about so-called "aspect retrieval" and "subtopic retrieval" are mostly close to our work. Here we regard the related work in terms of two categories: the interactive generality retrieval and the automatic generality ranking. The former is about how the generality is concerned in an interactive IR process while the latter is about how generality is associated in a ranking process.

2.1 Interactive Generality Retrieval

Interactive generality retrieval, or so-called "aspect retrieval", is studied in the interactive track of TREC-6,7,8 (Swan & Allan 1998, Robertson, Walker & Beaulieu 1999, Hersh 2000). The purpose of these studies is to help user retrieve documents covering as many different aspects of a topic as possible in a limited time. An aspect is defined as one of the many possible answers to the topic (Over 1999). Aspects of topics and documents in the collection are defined and judged by human assessors in order to evaluate the performance of aspect retrieval. In the area of aspect retrieval, researches are mainly focused on user's searching behavior and the interface of retrieval system.

We regard the aspect retrieval problem as a simplified version of the generality retrieval problem since the intuition is that the more aspects broadly covered by a document, the more general the document is. However, generality is richer than the aspect retrieval. Generality implies not only the broadness but also the deepness of a document in its meaning. Therefore our research will broaden the aspect retrieval into a problem of generality.

Furthermore, in order to help user's generality retrieval, automatic methods need to be developed. Given a query, automatic generality rankings is a process of ranking the retrieved documents by systematically estimating their generality. In the case of a large number of documents returned for a query, it is insufficient to improve the efficiency of the generality retrieval by improving the interface between user and the retrieval system. Automatic generality retrieval may be more efficient to help user to sort out documents with the consideration of generality. In next subsections, some researches closely related to automatic generality ranking are discussed.

2.2 Automatic Generality Ranking

Studies concerning automatic generality ranking aim at finding approaches to automatically rank general documents more closely to a query.

The study of subtopic retrieval (Zhai, Cohen & Lafferty 2003) seeks an automatic solution for the aspect retrieval problem we mentioned above. Zhai et al. 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" (Zhai et al. 2003). Given a set of documents retrieved by a baseline IR system, subtopic retrieval method re-ranks those documents by their generality feature and their relevance to the query. Statistical language models and maximal marginal relevance (Carbonell & Goldstain 1998) were used to perform subtopic retrieval.

Another research (Liu, Zhang, Chen, Lyu & Ma 2004) namely "affinity rank" is close to the study of subtopic retrieval. Affinity rank is based on the assumption that in a vector space model, "the more

neighbors a document has, the more informative it is; moreover, the more informative a document's neighbors are, the more informative it is as well" (Liu et al. 2004). Information richness was modeled by computing the principal eigenvector of a matrix M where each entry represents the value of a similarity function of each pair of documents in the vector space model.

The common feature of affinity ranking research and subtopic retrieval study is that document generality is based on the overall statistical properties of document in the collection rather than the concept generality. Concept generality is defined in our work as the generality of individual terms in the context of a given ontology. In WordNet, for example, hypernyms are defined as those concepts being more general than others; hyponyms are defined as those concepts being more specific than others. Since documents are composed of terms, document generality is consequently affected by the concept generality of all its terms.

Allen and Wu (Allen & Wu 2002) defined document generality as the mean generality of terms in the documents. For example, 64 selected words were determined manually as a reference collection for computing the generality. Half of the words in the collection were regarded as general and the other half as concrete. The 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 in those 64 terms of the reference collection, the generality of the terms in documents could be calculated.

However, some problems still remain unsolved. First, the generality of the terms in the reference list is determined by human experts. This is computationally infeasible to deal with a large number of words. Particularly if a term does not appear in the reference list, it is excluded from the generality computation. This is impractical in many applications that have a large vocabulary. It is expected that automatic methods can be developed to measure the concept generality objectively and efficiently for documents with a large domain-specific vocabulary. Secondly, not only the statistical term relatedness, but also the semantic relations between terms should be taken into account. Sometimes general terms may have low relatedness if they cross different domains. In the area of biomedical 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 at all, simply due to no co-occurrence in the text corpus. Third, in (Allen & Wu 2002), the generality was ranked for merely six documents and then manually judged for the evaluation. For dealing with large collections, this is obviously impractical. Finally, user generally would not prefer a document with high generality but low relevance to the query. Combining document generality with query generality should be considered. In next subsection, we briefly review some studies related to query generality.

2.3 Query Generality

To our knowledge, no researches considering query generality in document ranking or re-ranking process are currently available. Some definitions (He & Ounis 2004), (Plachouras, Cacheda, Ounis & Rijsbergen 2003), (Van Rijsbergen 1979) about query generality have been made long before the studies of document ranking. They mainly focus on the overall generality of retrieval rather than the generality of individual documents against a query. Van Rijsbergen (Van Rijsbergen 1979), (Plachouras et al. 2003) regarded query generality as "a measure of the density of relevant documents in the collection". Derived from Van Rijsbergen's definition, He and Ounis (He & Ounis 2004) defined query generality as:

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

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.

Based on these definitions of query generality, the more documents a query is related, the more generality the query has.

However, it is not sufficient to quantify the query generality purely based on this method. Let's consider two queries Q_1 "AIDS review" and Q_2 "SARS review". Q_1 requires literature reviews about AIDS, T_2 requires reviews about SARS, a newly discovered disease. In PubMed, Q_1 may result 19,311 documents. Whereas, there are only 396 documents returned by Q_2. Since it is hard to count the exact size of whole PubMed database, we assume that N is 11,000,000. According to Equation 1 the generality of Q_1 is around 6.3450. The generality of Q_2 is around 10.2320. Is Q_2 more general than Q_1? The answer is probably "no", because "SARS" is a newly discovered disease which has just less related documents in the collection than "AIDS".

In conclusion, there are some major differences between existing related work in the literature and our proposed approach.

1. We assume that the relevance judgment of a document is independent to that of the others retrieved by the same query. In the study of subtopic retrieval, relevance between two documents may depend on which documents a user sees the first.
2. We broaden the research problems of aspect retrieval, subtopic retrieval and affinity rank and propose the concept "generality" in document ranking.
3. Semantics inherence in the documents is considered in our research. We measure the ontology based semantic relationships of document concepts in order to compute generality. In literature, only statistical methods were used.
4. We consider both document generality and query generality. The documents are re-ranked by a combined score of similarity and the closeness of documents' generality to the query's. In literature, only document generality was considered.

In next section, we introduce the details of our ideas on re-ranking by generality.

3 Different Types of Generality

In our research, we divide generality into two categories based on user's information needs: technical generality and non-technical generality.

Technical Generality. How broad a document is for describing a certain topic. Documents with high technical generality are divided into two subcategories:

1. Summary
2. Review

Non-technical Generality. How deep a document is for describing a certain topic. Documents with high non-technical generality are divided into two categories:

1. Introduction
2. Tutorial

Technical generality should be considered when there is a need to retrieve summaries and technical review articles which broadly describe a certain topic. Non-technical generality should be considered when there is a need to retrieve introductive documents or tutorials that are jargon free and easy to understand. In this paper, we mainly focus on the study of technical generality, that is, on how to measure the document generality according to its broadness.

4 Proposed Approach

The intuition of our proposed computational generality is given as follows:

- *Document Scope* (DS) - We consider document as a collection of terms. The scope of a document is regarded as a coverage of terms onto the concepts in MeSH ontology. The more concepts matched within the MeSH the more specific the document is. Also, within a MeSH tree, the deeper the concepts appear, the more specific the document is.
- *Document Cohesion* (DC) - 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.

We formulate the problem of document ranking with generality as that: given a query Q , a rank $(R, \leq)$, $R = \{d_1, \ldots d_n\}$ which is retrieved by Q, the similarity function $Sim(Q, d_i)$,

1. find a function $Gen(Q, d_i)$ to return the closeness of generality between d_i and Q
2. re-rank $(R, \leq)$ to $(R', \leq)$ so that $d_i \leq d_j \iff f'(Sim(Q, d_i), Gen(Q, d_i)) \leq f'(Sim(Q, d_j), Gen(Q, d_j))$ where $d_i, d_j \in R'$. f' is a function considering both $Sim(Q, d_i)$ and $Gen(Q, d_i)$.

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 consideration is the semantics in documents. For instance, "HIV" is more specific than "virus" in terms of a given domain knowledge. The statistical analysis cannot reflect the semantic relationship between them.

A query can be regarded as a short document. In the same way, a query is to be computed for its generality as though it were 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.

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.

4.1 Ontology: MeSH Hierarchical Structure

All the headings used to index OHSUMED (Hersh et al. 1994) documents are well organized in a hierarchical structure namely MeSH tree. Figure 1 is a fragment of the MeSH tree.

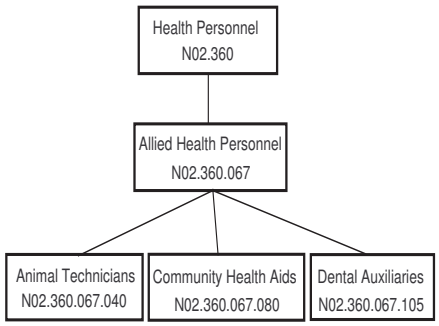

Figure 1: A Fragment of MeSH tree

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 "Allied 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 the 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.

4.2 Computation of Document Generality

4.2.1 Concept Identification

In order to use MeSH ontology to extract the semantic relations between terms, the MeSH concepts in the text corpus must be recognized. An algorithm of concept identification is proposed to match single or compound(noun) terms in the corpus with the concepts in the MeSH tree.

The algorithm is mainly concerned with the subsumed terms: 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 out of 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 is 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.

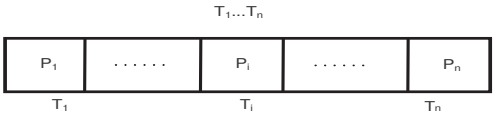

Figure 2: Data Structure of a Node in CMT

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 goto 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. [2]
```

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 no other candidate concepts can be found with more compound terms than concept C in the same place of the document, 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.

4.2.2 Computing Document Scope

Document scope is about how broad or vague a document is for describing a certain topic. It is an important feature of document generality. Consider the following two definitions of SARS. Definition 1 comes from ABOUT [3], a web information service for daily life. Definition 2 is an official definition from the Department of Health in Hong Kong [4].

1. A viral respiratory illness that was recognized as a global threat in March 2003.
2. A viral respiratory infection caused by a coronavirus (SARS-CoV).

In above definition 2 we may identify three MeSH concepts: "respiratory infection", "coronavirus" and "SARS-CoV". However, in definition 1 which is for the general public, no MeSH concept is found. "Respiratory illness" is used to broadly describe SARS rather than a more narrowed concept "respiratory infection".

We mentioned that the scope of a document is regarded as a coverage of terms onto the concepts in MeSH ontology. The more concepts matched within the MeSH the more specific the document is. Also, within a MeSH tree, the deeper the concepts appear, the more specific the document is. In our computation of document scope, both MeSH concepts and non-ontology concepts in document are considered. Firstly a mean function of tree depths of all concepts in document is proposed to calculate document scope. The depth of a MeSH concept is measured by the distance between that concept and the root of the MeSH tree. The tree depth of a non-ontology concept in MeSH tree is zero. Secondly, we normalize the scope function within the range of 0 and 1.

It is often the case that a document contains a large percentage of non-ontology concepts but just a small percentage of MeSH concepts. This kind of documents may have a low average tree depth of all concepts and may be close to each other in terms of their computed scope values. Therefore, we need to make the scope function to be more sensitive to documents with low average tree depths compared with that of the documents with high average tree depths. In our research, we select an exponential function that can well satisfy our requirement for the distribution of scope function values.

$$Scope(d_i) = e^{-\left(\frac{\sum_{i=1}^{n} depth(c_i)}{n}\right)} \qquad (2)$$

In Equation 2, n is the total number of concepts of both MeSH concepts and general concepts. Moreover, stop words are excluded in this example. Function $depth(c_i)$ is to get the tree depth of concept i in the MeSH tree. As to a document which contains only non-ontology concepts, its document scope is 1, the maximum value. For a document which has maximum average tree depth of all its MeSH concepts, its scope is e^{-11}, the minimum value. The time complexity of scope-based ranking is $O(m \times n)$, m is the number of retrieved documents, n is the average concepts in those documents.

There are two typical examples where the concepts in documents may have different distributions in MeSH tree in terms of their subsumption relationships. Concept A subsuming concept B in the MeSH tree indicates that A is one of the parent nodes of B. The followings are illustrations of our scope algorithm in both examples.

Example One

A document may contain MeSH concepts that have no subsumption relationship between each

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

[3] http://about.com

[4] http://www.info.gov.hk

other in the MeSH tree. In Figure 3, there is a piece of MeSH tree. Every labeled node is a MeSH concept. Suppose that d_i and d_j are two documents in the document collection. d_i is more general than d_j. Each of them contains only two concepts. The concepts o and p in d_i have matches found in the MeSH tree (the darkened nodes). The concepts k and h in d_j have matches found in the MeSH tree too. According to our algorithm, the average tree depths of d_i and d_j are respectively 3 and 4. The scope of d_i is 0.0498, which is greater than 0.0183, the scope of d_j.

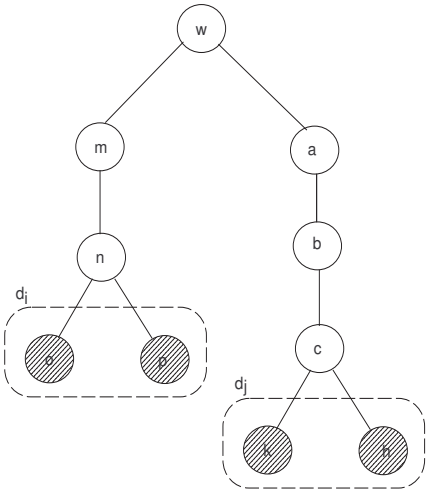

Figure 3: d_i and d_j with different document scope

Example Two

A document may contain MeSH concepts that have subsumption relationship between each other in the MeSH tree. In Figure 4, there is a piece of MeSH tree. Every labeled node is a MeSH concept. Suppose that d_i and d_j are two documents in the document collection. d_i is more general than d_j. Each of them contains only two concepts. The concepts m and n in d_i have matches found in the MeSH tree (the darkened nodes). The concepts c and h in d_j have matches found in the MeSH tree too. According to our algorithm, the average tree depths of d_i and d_j are respectively 1.5 and 3.5. The scope of d_i is 0.2231, which is greater than 0.0302, the scope of d_j.

In above SARS example, $S(d_1)$ is 1 because no MeSH concept can be found. As to $S(d_2)$, $Depth(\text{“}respiratory\ infection''\text{)} = 3$, $Depth(\text{“}coronavirus''\text{)} = 5.5$ since there are two nodes in MeSH tree representing “coronavirus”, one has a depth 5 and another is 6. An average tree depth is calculated in this example. $Depth(\text{“}SARS-CoV''\text{)} = 6.5$. The total number of concepts in definition 2 is 8. Therefore the value of $S(d_2)$ is 0.1534 which is smaller than $S(d_1)$. This result shows definition 2 has less generality than definition 1.

4.2.3 Computing Document Cohesion

With MeSH hierarchical structure (tree), it is possible to retrieve the semantic distance between MeSH concepts according to their positions in the 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” (Morris & Hirst 1991). The intuition of our approach is based on a hypothesis that document with less cohesion would be more general. Consider two definitions of HIV: the first one comes from a web site called AIDS 101, Guide to HIV basics[5], and the second come from MeSH ontology. Obviously, definition 1 is more general than definition 2.

1. “HIV-1” is the virus most researchers believe causes AIDS.
2. HIV is a non-taxonomic and historical term referring to any of two species, specifically HIV-1 and/or HIV-2.

In definition 2, three MeSH concepts can be identified: “HIV”, “HIV-1” and “HIV-2”. In definition 1, “HIV”, “AIDS” and “virus” are identified as MeSH concepts.

What causes definition 1 to 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. “HIV-1” and “HIV-2” are two types of “HIV” in terms of MeSH ontology. However, in definition 1, “HIV” is a kind of virus but “AIDS” is a kind of diseases. There is not a direct relationship between them. Moreover, “HIV” doesn't directly belong to “virus” in MeSH tree.

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.

We mentioned that the cohesion of a document is regarded as a computation of the associations between the concepts found in the MeSH tree. The more closely the concepts are associated, the more specific the document is. In terms of that, firstly, in our computation of document scope MeSH concepts in document are considered rather than non-ontology concepts. A mean function is used to calculate the average strength of associations between all pairs of MeSH concepts found in document. Secondly, we assume that the strength of association between two MeSH concepts is a monotonic decreasing function of the shortest path between them in the MeSH tree. The minimum value of the function is set to 0 when the shortest path between two MeSH concepts is as large as twice the maximum tree depth. The maximum value of the function is resulted when the shortest path between them equals to 1. In our

[5] http://www.sfaf.org/aids101/

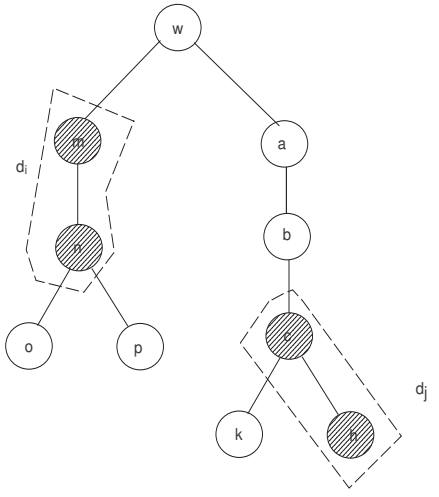

Figure 4: d_i and d_j with different document scope

research, the calculation of semantic association between concepts is based on the Leacock-Chodorow similarity (Leacock & Chodorow 1998) function which is a logistic function featured for measuring the shortest path between two concepts in the MeSH tree.

$$Cohesion(d_i) = \frac{\sum_{i,j=1}^{n} Sim(c_i, c_j)}{Numberof Associations},\ (n > 1,\ i < j) \quad (3)$$

$$Sim(c_i, c_j) = -log\frac{len(c_i, c_j)}{2D} \quad (4)$$

$$Numberof Associations = \frac{n(n-1)}{2} \quad (5)$$

In Equation 3, 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 5.

In Equation 4, D is the maximum MeSH tree depth. In our experiments, D is 11. The scope of Equation 3 is $[0, -log(\frac{1}{22})]$. As to a document with zero or one MeSH concept only, its document cohesion is set to 0. For a documents with strongest associations among all the concepts within the document, its cohesion is $-log(\frac{1}{22})$, the maximum value. The time complexity of cohesion-based ranking is $O(m \times n^2)$, m is the number of retrieved documents, n is the average concepts in those documents.

There are two typical examples where the concepts in documents may have different distributions in MeSH tree in terms of their subsumption relationships. Concept A subsuming concept B in the MeSH tree indicates that A is one of the parent nodes of B. The followings are illustrations of our cohesion algorithm in both examples.

Example Three

A document may contain MeSH concepts that have no subsumption relationship between each other in the MeSH tree. In Figure 5, there is a piece of MeSH tree. Every labeled node is a MeSH concept. Suppose that d_i and d_j are two documents in the document collection. d_j is more general than d_i. Each of them contains only two concepts. The concepts o and p in d_i have matches found in the MeSH tree (the darkened nodes). The concepts x and y in d_j have matches found in the MeSH tree too. According to our algorithm, the length of the shortest path between o and p is 2. The shortest distance between x and y is 4. Thus the cohesion of d_i is 2.3979, greater than the generality of d_j, 1.7047.

Example Four

A document may contain MeSH concepts that have subsumption relationship between each other in the MeSH tree. In Figure 6, there is a piece of MeSH tree. Every labeled node is a MeSH concept. Suppose that d_i and d_j are two documents in the document collection. d_j is more general than d_i. Each of them contains only two concepts. The concepts o and n in d_i have matches found in the MeSH tree (the darkened nodes). The concepts i and y in d_j have matches found in the MeSH tree too. According to our algorithm, the length of the shortest path between o and n is 1. The shortest distance between i and y is 2. Thus the cohesion of d_i is 3.0910, greater than the generality of d_j, 2.3979.

4.2.4 Computing Document Generality

The following is the formula for the calculation of document generality.

$$DG(d_i) = \frac{Scope(d_i)}{Cohesion(d_i) + 1} \quad (6)$$

The query generality computation is similar to the computation of document generality. The difference between them is that we take ω, the Statistical Query Generality (SQG), in Equation 1 as an optional parameter for query generality calculation.

$$QG = \frac{SQG * Scope(Q)}{Cohesion(Q) + 1} \quad (7)$$

In Equation 7, QG is the query generality. The calculations of query cohesion and scope is the same as document cohesion and scope.

However, we argue 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.

4.2.5 Correlation Analysis

The independent relationship between generality and similarity is a major hypothesis of this paper. Prov-

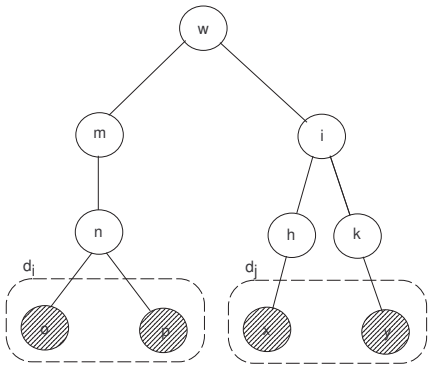

Figure 5: d_i and d_j with different document cohesion

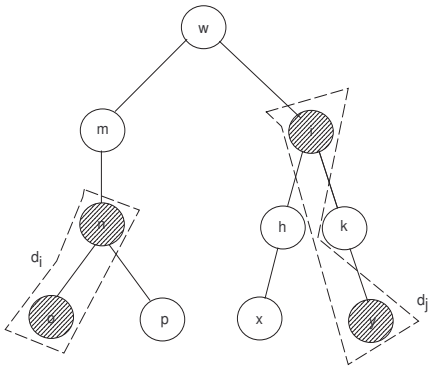

Figure 6: d_i and d_j with different document cohesion

ing this hypothesis is important for the further combination of similarity and generality in the re-ranking process.

Theoretically, the similarity computation itself, does not reflect the generality that exists within the documents. For example, we assume that user can explicitly specify for a given query, if the query is intended to be specific, or general, then the conventional similarity-based methods would not be able to retrieve documents which are specific or general in a given domain knowledge.

Practically, we clarify the relationship between generality and similarity by using intuitive scatter diagrams and Pearson's correlation coefficients. Pearson's correlation coefficient can measure the degree of association between two continuous variables. Scatter diagram can visualize their association. By submitting query to a IR system, the similarity of retrieved documents to query are computed and scored. Our proposed method to calculate document generality is then applied on those retrieved documents in order to get their generality score. The correlation analysis between document generality and its similarity score is then used to prove our hypothesis of the independent relationship between generality and similarity.

4.2.6 Combining Similarity and Generality

As an important step in our proposed approach, we consider both the document similarity and generality. Here information retrieval system is regarded 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 similarity score given to a ranked document d_i and QG is the query generality. The final score considering both document similarity and generality is given in the following formula.

$$Score(d_i, Q) = RScore(d_i)^{\alpha} * e^{-|DG(d_i)-QG|^{\beta}} \quad (8)$$

α and β are parameters for a well tuned performance.

5 Experiment and Evaluation

It is necessary to evaluate the effectiveness of our proposed algorithm. The evaluation of effectiveness can be divided into two aspects. The first is on how it can improve the overall performance of a baseline IR system, while the second is on how it can improve the performance of the generality retrieval.

We evaluated the effectiveness of our proposed re-ranking algorithm on the overall query performance by comparing our algorithm against a baseline IR system.

5.1 Data Set and Queries

Our model has been evaluated on the OHSUMED (Hersh et al. 1994) 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 OHSUMED (Hersh et al. 1994) 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 because 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

5.2 Baseline and Pre-processing

Lucene[6] is used as the baseline IR system to index and retrieve the titles and abstracts of documents in OHSUMED collection (Hersh et al. 1994). We chose Lucene as our baseline IR system as it offers a full representative features of a traditional keyword matching IR system. 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.

5.3 Evaluation Methodology

In our experiments, the baseline IR system is used to retrieve 1000 documents for each test query. We then cover all nine possible cases where query generality, document generality and SQG are used solely or together in a reasonable manner. Those nine cases are derived from our proposed Equation 6, 7 and 8 for re-ranking the documents retrieved by the baseline IR system. For example, DS is the case where only document scope (i.e. Equation 2) is considered in the computation of document generality. The score function in Equation 8 is then simplified as Equation 9 and 10 where α and β are parameters for a well tuned performance.

$$DG(d_i) = Scope(d_i) \quad (9)$$

$$Score(d_i, Q) = RScore(d_i)^{\alpha} * DG(d_i)^{\beta} \quad (10)$$

QS+QC+DS+DC is the case where the closeness between query generality (scope and cohesion) and document generality (scope and cohesion) is considered in the computation of generality re-ranking(i.e. Equation 8). QS and QC denote query scope and query cohesion, DS and DC denote document scope and document cohesion.

5.4 Performance Indicators

The performance of re-ranking is measured in two aspects. Firstly we compare the precision and recall of re-ranking with the original ranking given by baseline IR system[7] for all the 101 test queries. Secondly, we check if all the review type queries get larger improvement in term of average precision.

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

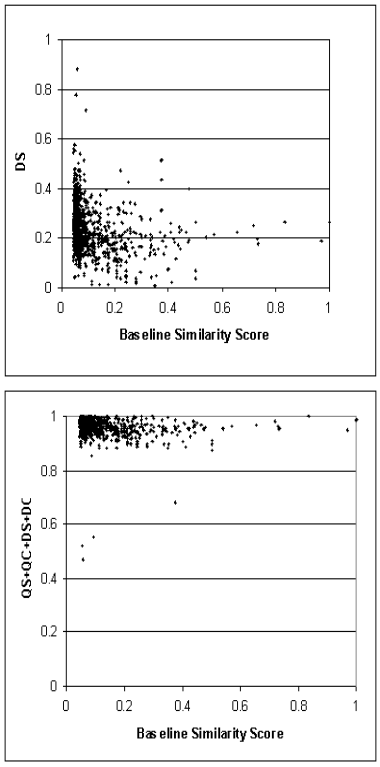

Figure 7: (Query No.5) Up: DS R = -0.24 Down: QS+QC+DS+DC R = -0.17

5.5 Experiment Results

In the upper part of Figure 7, the correlation between DS case and the baseline IR system is shown in a scatter diagram. In the lower part of Figure 7, the correlation between QS+QC+DS+DC case and the baseline IR system is shown in a scatter diagram.

Figure 8 shows the precision-recall graph in a certain range of precision and recall. Due to the limitation of space, Table 1 shows only the detailed precisions of one of the nine cases with the best performance at different recall levels. In Table 2, we show how the review type queries are improved by a comparison of mean average precision between our proposed re-ranking algorithms and the baseline IR system. The mean average precision ("MAP" in the tables) and the percentages of improvement in MAP ("%" in the tables) are summarized.

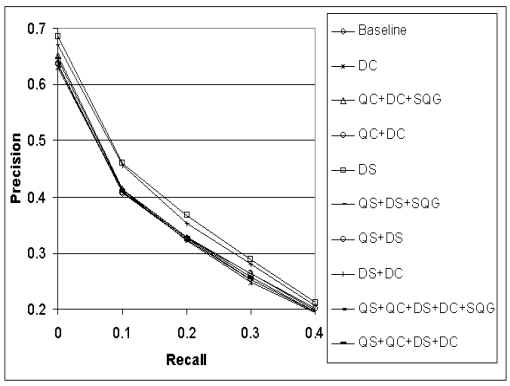

Figure 8: Precision Recall Graph of Overall Query Performance (Recall in [0, 0.4], Precision in [0.2, 0.7])

Table 1: Detailed Precision-Recall Comparisons

Recall	Baseline	DS
0	0.6369	**0.6858**
0.1	0.4071	**0.4591**
0.2	0.3239	**0.3674**
0.3	0.254	**0.2881**
0.4	0.1963	**0.2125**
0.5	0.1679	**0.1770**
0.6	0.1396	**0.1414**
0.7	0.088	**0.0917**
0.8	0.0544	**0.0565**
0.9	0.0223	**0.0236**
1	0.0018	**0.0023**
MAP	0.1849	**0.2036**
%		**10.11%**
R-prec	0.2246	**0.2800**
%		**24.67%**

Table 2: Precision Improvement on Review Type Queries

QNo.	Baseline	QS+QC+DS+DC
4	0.0821	**0.0827**
11	0.0741	**0.0935**
17	0.0021	**0.0023**
31	0.1522	**0.1525**
34	0.0193	0.0190
54	0.1099	**0.1124**
105	0.2950	0.2949
106	0.0085	**0.0087**
MAP	0.0929	**0.0958**
%		**3.07%**

5.6 Results Analysis

In Figure 7, it can be clearly seen that there is no strong relationship (e.g. linear relationship) between generality and similarity in the scatter diagrams. Moreover, the values of correlation coefficient are quite small too. Therefore it shows that generality and similarity are two different concepts without strong correlation between them.

Within all the cases, DS improve the query performance significantly for all 101 queries. There are a 10.11% improvement of MAP and 24.67% improvement of R-prec. This indicates that it is effective to do the re-ranking by considering both document generality and similarity.

The results show the better performance of QS+QC+DS+DC on review type queries. There is an encouraging 3.07% improvement for QS+QC+DS+DC. We performed a dependent t-test (Paired Two Sample for Means) which compares the paired precisions between the baseline and the QS+QC+DS+DC 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 that the technical generality retrieval happens more often for review type queries from non-domain-expert user.

6 Conclusions

In this paper, we argued that there is a need of document generality computation in information retrieval. A novel approach to generality computation has been proposed. Our approach uses the MeSH ontology structure in bio-medical domain to compute the gen-

erality based on both statistical and semantic relationships between the terms. Then we applied our proposed generality computation method to the document re-ranking in bio-medical information retrieval. Traditional similarity-based document ranking methods are incorporated with the generality computations. The experiments of our approach have shown that "generality" is an important complement to the traditional similarity-based ranking. The intuition is that when search results are returned by IR system, user may expect to see the documents broadly describing a certain topic to be ranked on the top of the list, so that they can get an overview of the topic first rather than going into the specific ones immediately.

In our proposed framework of document re-ranking in bio-medical information retrieval, documents are scored and re-ranked by a combination of their similarity to query and the closeness of documents' generality to the query's. Experiments have been conducted on a large corpus namely OHSUMED (Hersh et al. 1994). Our approach shows an improved query performance and encourages us to pursue the further investigation. Our approach can also be applicable to other domains where the domain specific ontology is available.

There are some further works expected. Firstly the cohesion algorithm currently has an oversimplification since it considers semantic relationship between MeSH concepts only. Since there is a large percentage of non-ontology concepts in documents, it is necessary to consider statistical relationships between concepts. A possible solution is to consider the co-occurrence relationship of concepts (i.e. both MeSH and non-ontology concept). The more often two concepts co-occur, the stronger their association is.

Secondly, it is necessary to fully evaluate the effectiveness of our proposed algorithms on generality-based retrieval by comparing our algorithms with other baselines (Zhai et al. 2003), (Liu et al. 2004). More experiments on the evaluation frameworks in related work (Zhai et al. 2003), (Liu et al. 2004) need to be performed for the purpose of tuning the generality computation formulas.

Finally, the domain-independent generality ranking may need to be studied. Currently our proposed algorithms are domain dependent. We re-rank biomedical documents in the context of a given biomedical ontology. The performance of our re-ranking algorithms in a general domain-independent environment is unknown. However, the idea presented in this paper has shown a new way of document ranking and is promising towards the improvement of information retrieval in general.

7 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

Allen, R. B. & Wu, Y. (2002), Generality of texts, *in* 'Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology', pp. 111–116.

Carbonell, J. & Goldstain, J. (1998), The use of MMR, diversity-based reranking for reordering documents and producing summaries., *in* 'SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval', ACM Press, New York, NY, USA.

He, B. & Ounis, I. (2004), Inferring query performance using pre-retrieval predictors, *in* '11th Symposium on String Processing and Information Retrieval', Padova, Italy, pp. 43–54.

Hersh, W. (2000), TREC-8 interactive track report, *in* 'The Eighth Text REtrieval Conference', pp. 57–64.

Hersh, W., Buckley, C., Leone, T. J. & Hickam, D. (1994), OHSUMED: an interactive retrieval evaluation and new large test collection for research, *in* 'Annual ACM Conference on Research and Development in Information Retrieval', pp. 192 – 201.

Leacock, C. & Chodorow, M. (1998), Combining local context and wordnet similarity for word sense identification, *in* 'Fellbaum', pp. 265–283.

Liu, Y., Zhang, B., Chen, Z., Lyu, M. R. & Ma, W.-Y. (2004), Affinity rank: A new scheme for efficient web search, *in* 'The Thirteenth World Wide Web conference', Vol. 203-211, ACM, New York, USA.

Morris, J. & Hirst, G. (1991), 'Lexical cohesion computed by thesaural relations as an indicator of the structure of text', *Computational Linguistics* **17**(1), 21–48.

Over, P. (1999), TREC-7 interactive track report, *in* 'The Seventh Text REtrieval Conference', pp. 65–72.

Plachouras, V., Cacheda, F., Ounis, I. & Rijsbergen, C. v. (2003), 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.

Robertson, S. E., Walker, S. & Beaulieu, M. (1999), Okapi at TREC-7: automatic ad hoc, filtering, vlc and interactive track., *in* E. M. Voorhees & D. K. Harman, eds, 'The Seventh Text REtrieval Conference (TREC-7)', Gaithersburg, MD, USA.

Salton, G., Wong, A. & Yang, C. S. (1975), 'A vector space model for automatic indexing', *Communications of the ACM* **18**(11).

Swan, R. C. & Allan, J. (1998), Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems, *in* 'SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval', ACM Press, New York, NY, USA, pp. 173–181.

Van Rijsbergen, C. J. (1979), *Information Retrieval*, London; Boston: Butterworths.

Zhai, C., Cohen, W. W. & Lafferty, J. (2003), 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', pp. 10–17.