Knowledge-intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature

Wei Zhou, Clement Yu
Department of Computer Science
University of Illinois at Chicago
wzhou8@uic.edu,
yu@cs.uic.edu

Neil Smalheiser, Vetle Torvik
Department of Psychiatry and
Psychiatric Institute (MC912)
University of Illinois at Chicago
{neils, vtorvik}@uic.edu

Jie Hong
Division of Epidemiology and
Biostatistics, School of Public health
University of Illinois at Chicago
jhong20@uic.edu

ABSTRACT

This paper presents a study of incorporating domain-specific knowledge (i.e., information about concepts and relationships between concepts in a certain domain) in an information retrieval (IR) system to improve its effectiveness in retrieving biomedical literature. The effects of different types of domain-specific knowledge in performance contribution are examined. Based on the TREC platform, we show that appropriate use of domain-specific knowledge in a proposed conceptual retrieval model yields about 23% improvement over the best reported result in passage retrieval in the Genomics Track of TREC 2006.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models, query formulation, information filtering.* H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *thesauruses.*

General Terms

Algorithms, Performance, Experimentation.

Keywords

Document Retrieval, Passage Extraction, Biomedical Documents

1. INTRODUCTION

Biologists search for literature on a daily basis. For most biologists, PubMed, an online service of U.S. National Library of Medicine (NLM), is the most commonly used tool for searching the biomedical literature. PubMed allows for keyword search by using Boolean operators. For example, if one desires documents on the use of the drug propanolol in the disease hypertension, a typical PubMed query might be "propanolol AND hypertension", which will return all the documents having the two keywords. Keyword search in PubMed is effective if the query is well-crafted by the users using their expertise. However, information needs of biologists, in some cases, are expressed as complex questions [8][9], which PubMed is not designed to handle. While NLM does maintain an experimental tool for free-text queries [6], it is still based on PubMed keyword search.

The Genomics track of the 2006 Text REtrieval Conference (TREC) provides a common platform to assess the methods and techniques proposed by various groups for biomedical information retrieval. The queries were collected from real biologists and they are expressed as complex questions, such as "How do mutations in the Huntingtin gene affect Huntington's disease?". The document collection contains 162,259 Highwire full-text documents in HTML format. Systems from participating groups are expected to find relevant passages within the full-text documents. A passage is defined as any span of text that does not include the HTML paragraph tag (i.e., <P> or </P>).

We approached the problem by utilizing domain-specific knowledge in a conceptual retrieval model. Domain-specific knowledge, in this paper, refers to information about concepts and relationships between concepts in a certain domain. We assume that appropriate use of domain-specific knowledge might improve the effectiveness of retrieval. For example, given a query "What is the role of gene PRNP in the Mad Cow Disease?", expanding the gene symbol "PRNP" with its synonyms "Prp", "PrPSc", and "prion protein", more relevant documents might be retrieved. PubMed and many other biomedical systems [8][9][10][13] also make use of domain-specific knowledge to improve retrieval effectiveness.

Intuitively, retrieval on the level of concepts should outperform "bag-of-words" approaches, since the semantic relationships among words in a concept are utilized. In some recent studies [13][15], positive results have been reported for this hypothesis. In this paper, concepts are entry terms of the ontology Medical Subject Headings (MeSH), a controlled vocabulary maintained by NLM for indexing biomedical literature, or gene symbols in the Entrez gene database also from NLM. A concept could be a word, such as the gene symbol "PRNP", or a phrase, such as "Mad cow diseases". In the conceptual retrieval model presented in this paper, the similarity between a query and a document is measured on both concept and word levels.

This paper makes two contributions:

1. We propose a conceptual approach to utilize domain-specific knowledge in an IR system to improve its effectiveness in

SIGIR'07, July 23–27, 2007, Ámsterdam, The Netherlands.
Copyright 2007 ACM 978-1-59593-597-7/07/0007...$5.00.

retrieving biomedical literature. Based on this approach, our system achieved significant improvement (23%) over the best reported result in passage retrieval in the Genomics track of TREC 2006.

2. We examine the effects of utilizing concepts and of different types of domain-specific knowledge in performance contribution.

This paper is organized as follows: problem statement is given in the next section. The techniques are introduced in section 3. In section 4, we present the experimental results. Related works are given in section 5 and finally, we conclude the paper in section 6.

2. PROBLEM STATEMENT

We describe the queries, document collection and the system output in this section.

The query set used in the Genomics track of TREC 2006 consists of 28 questions collected from real biologists. As described in [8], these questions all have the following general format:

$$\text{Biological object (1..m)} \xleftrightarrow{\text{Relationship}} \text{Biological process (1..n)} \qquad (1)$$

where a biological object might be a gene, protein, or gene mutation and a biological process can be a physiological process or disease. A question might involve multiple biological objects (m) and multiple biological processes (n). These questions were derived from four templates (Table 2).

Table 2 Query templates and examples in the Genomics track of TREC 2006

Template	Example
What is the role of gene in disease?	What is the role of DRD4 in alcoholism?
What effect does gene have on biological process?	What effect does the insulin receptor gene have on tumorigenesis?
How do genes interact in organ function?	How do HMG and HMGB1 interact in hepatitis?
How does a mutation in gene influence biological process?	How does a mutation in Ret influence thyroid function?

Features of the queries: 1) They are different from the typical Web queries and the PubMed queries, both of which usually consist of 1 to 3 keywords; 2) They are generated from structural templates which can be used by a system to identify the query components, the biological object or process.

The document collection contains 162,259 Highwire full-text documents in HTML format.

The output of the system is a list of passages ranked according to their similarities with the query. A passage is defined as any span of text that does not include the HTML paragraph tag (i.e., <P> or </P>). A passage could be a part of a sentence, a sentence, a set of consecutive sentences or a paragraph (i.e., the whole span of text that are inside of <P> and </P> HTML tags).

This is a passage-level information retrieval problem with the attempt to put biologists in contexts where relevant information is provided.

3. TECHNIQUES AND METHODS

We approached the problem by first retrieving the top-*k* most relevant paragraphs, then extracting passages from these paragraphs, and finally ranking the passages. In this process, we employed several techniques and methods, which will be introduced in this section. First, we give two definitions:

Definition 3.1 A **concept** is 1) a entry term in the MeSH ontology, or 2) a gene symbol in the Entrez gene database. This definition of concept can be generalized to include other biomedical dictionary terms.

Definition 3.2 A **semantic type** is a category defined in the Semantic Network of the Unified Medical Language System (UMLS) [14]. The current release of the UMLS Semantic Network contains 135 semantic types such as “Disease or Syndrome”. Each entry term in the MeSH ontology is assigned one or more semantic types. Each gene symbol in the Entrez gene database maps to the semantic type “Gene or Genome”. In addition, these semantic types are linked by 54 relationships. For example, “Antibiotic” *prevents* “Disease or Syndrome”. These relationships among semantic types represent general biomedical knowledge. We utilized these semantic types and their relationships to identify related concepts.

The rest of this section is organized as follows: in section 3.1, we explain how the concepts are identified within a query. In section 3.2, we specify five different types of domain-specific knowledge and introduce how they are compiled. In section 3.3, we present our conceptual IR model. Finally, our strategy for passage extraction is described in section 3.4.

3.1 Identifying concepts within a query

A concept, defined in **Definition 3.1**, is a gene symbol or a MeSH term. We make use of the query templates to identify gene symbols. For example, the query “How do HMG and HMGB1 interact in hepatitis?” is derived from the template “How do genes interact in organ function?”. In this case, “HMG” and “HMGB1” will be identified as gene symbols. In cases where the query templates are not provided, programs for recognition of gene symbols within texts are needed.

We use the query translation functionality of PubMed to extract MeSH terms in a query. This is done by submitting the whole query to PubMed, which will then return a file in which the MeSH terms in the query are labeled. In Table 3.1, three MeSH terms within the query “What is the role of gene PRNP in the Mad cow disease?” are found in the PubMed translation: "encephalopathy, bovine spongiform" for “Mad cow disease”, “genes” for “gene”, and “role” for “role”.

Table 3.1 The PubMed translation of the query "What is the role of gene PRNP in the Mad cow disease?".

Term	PubMed translation
Mad cow disease	"bovine spongiform encephalopathy"[Text Word] OR "encephalopathy, bovine spongiform"[MeSH Terms] OR Mad cow disease[Text Word]
gene	("genes"[TIAB] NOT Medline[SB]) OR "genes"[MeSH Terms] OR gene[Text Word]
role	"role"[MeSH Terms] OR role[Text Word]

3.2 Compiling domain-specific knowledge

In this paper, domain-specific knowledge refers to information about concepts and their relationships in a certain domain. We used five types of domain-specific knowledge in the domain of genomics:

Type 1. Synonyms (terms listed in the thesauruses that refer to the same meaning)
Type 2. Hypernyms (more generic terms, one level only)
Type 3. Hyponyms (more specific terms, one level only)
Type 4. Lexical variants (different forms of the same concept, such as abbreviations. They are commonly used in the literature, but might not be listed in the thesauruses)
Type 5. Implicitly related concepts (terms that are semantically related and also co-occur more frequently than being independent in the biomedical texts)

Knowledge of type 1-3 is retrieved from the following two thesauruses: 1) MeSH, a controlled vocabulary maintained by NLM for indexing biomedical literature. The 2007 version of MeSH contains information about 190,000 concepts. These concepts are organized in a tree hierarchy; 2) Entrez Gene, one of the most widely used searchable databases of genes. The current version of Entrez Gene contains information about 1.7 million genes. It does not have a hierarchy. Only synonyms are retrieved from Entrez Gene. The compiling of type 4-5 knowledge is introduced in section 3.2.1 and 3.2.2, respectively.

3.2.1 Lexical variants

Lexical variants of gene symbols

New gene symbols and their lexical variants are regularly introduced into the biomedical literature [7]. However, many reference databases, such as UMLS and Entrez Gene, may not be able to keep track of all this kind of variants. For example, for the gene symbol "NF-kappa B", at least 5 different lexical variants can be found in the biomedical literature: "NF-kappaB", "NFkappaB", "NFkappa B", "NF-kB", and "NFkB", three of which are not in the current UMLS and two not in the Entrez Gene. [3][21] have shown that expanding gene symbols with their lexical variants improved the retrieval effectiveness of their biomedical IR systems. In our system, we employed the following two strategies to retrieve lexical variants of gene symbols.

Strategy I: This strategy is to automatically generate lexical variants according to a set of manually crafted heuristics [3][21]. For example, given a gene symbol “PLA2”, a variant “PLAII” is generated according to the heuristic that Roman numerals and Arabic numerals are convertible when naming gene symbols. Another variant, “PLA 2”, is also generated since a hyphen or a space could be inserted at the transition between alphabetic and numerical characters in a gene symbol.

Strategy II: This strategy is to retrieve lexical variants from an abbreviation database. ADAM [22] is an abbreviation database which covers frequently used abbreviations and their definitions (or long-forms) within MEDLINE, the authoritative repository of citations from the biomedical literature maintained by the NLM. Given a query “How does nucleoside diphosphate kinase (NM23) contribute to tumor progression?”, we first identify the abbreviation “NM23” and its long-form “nucleoside diphosphate kinase” using the abbreviation identification program from [4]. Searching the long-form “nucleoside diphosphate kinase” in ADAM, other abbreviations, such as “NDPK” or “NDK”, are retrieved. These abbreviations are considered as the lexical variants of “NM23”.

Lexical variants of MeSH concepts

ADAM is used to obtain the lexical variants of MeSH concepts as well. All the abbreviations of a MeSH concept in ADAM are considered as lexical variants to each other. In addition, those long-forms that share the same abbreviation with the MeSH concept and are different by an edit distance of 1 or 2 are also considered as its lexical variants. As an example, "human papilloma viruses" and "human papillomaviruses" have the same abbreviation “HPV” in ADAM and their edit distance is 1. Thus they are considered as lexical variants to each other. The edit distance between two strings is measured by the minimum number of insertions, deletions, and substitutions of a single character required to transform one string into the other [12].

3.2.2 Implicitly related concepts

Motivation: In some cases, there are few documents in the literature that directly answer a given query. In this situation, those documents that implicitly answer their questions or provide supporting information would be very helpful. For example, there are few documents in PubMed that directly answer the query "What is the role of the genes HNF4 and COUP-tf I in the suppression in the function of the liver?". However, there exist some documents about the role of "HNF4" and "COUP-tf I" in regulating "hepatitis B virus" transcription. It is very likely that the biologists would be interested in these documents because "hepatitis B virus" is known as a virus that could cause serious damage to the function of liver. In the given example, "hepatitis B virus" is not a synonym, hypernym, hyponym, nor a lexical variant of any of the query concepts, but it is semantically related to the query concepts according to the UMLS Semantic Network. We call this type of concepts “implicitly related concepts” of the query. This notion is similar to the “*B*-term” used in [19] for relating two disjoint literatures for biomedical hypothesis generation. The difference is that we utilize the semantic relationships among query concepts to exclusively focus on concepts of certain semantic types.

A query q in format (1) of section 2 can be represented by

$$q = (A, C)$$

where A is the set of biological objects and C is the set of biological processes. Those concepts that are semantically related to both A and C according to the UMLS Semantic Network are considered as the implicitly related concepts of the query. In the above example, A = {“HNF4”, “COUP-tf I”}, C = {“function of liver”}, and "hepatitis B virus" is one of the implicitly related concepts.

We make use of the MEDLINE database to extract the implicitly related concepts. The 2006 version of MEDLINE database contains citations (i.e., abstracts, titles, and etc.) of over 15 million biomedical articles. Each document in MEDLINE is manually indexed by a list of MeSH terms to describe the topics covered by that document. Implicitly related concepts are extracted and ranked in the following steps:

Step 1. Let list_A be the set of MeSH terms that are 1) used for indexing those MEDLINE citations having A, and 2) semantically related to A according to the UMLS Semantic Network. Similarly, list_C is created for C. Concepts in B = list_$A \cap$ list_C are considered as implicitly related concepts of the query.

Step 2. For each concept $b \in B$, compute the association between *b* and A using the mutual information measure [5]:

$$I(b,A) = \log \frac{\mathrm{P}(b,A)}{\mathrm{P}(b)\mathrm{P}(A)}$$

where $\mathrm{P}(x) = n/N$, *n* is the number of MEDLINE citations having *x* and *N* is the size of MEDLINE. A large value for *I*(*b, A*) means that *b* and *A* co-occur much more often than being independent. *I*(*b, C*) is computed similarly.

Step 3. Let $r(b) = (I(b, A), I(b, C))$, for $b \in B$. Given $b_1, b_2 \in B$, we say $r(b_1) \leq r(b_2)$ if $I(b_1, A) \leq I(b_2, A)$ and $I(b_1, \mathrm{C}) \leq I(b_2, C)$. Then the association between *b* and the query *q* is measured by:

$$score(b,q) = \frac{\left|\{x : x \in B \text{ and } r(x) \leq r(b)\}\right|}{\left|\{x : x \in B \text{ and } r(b) \leq r(x)\}\right|} \quad (2)$$

The numerator in Formula 2 is the number of the concepts in *B* that are associated with both *A* and *C* equally with or less than *b*. The denominator is the number of the concepts in *B* that are associated with both *A* and *C* equally with or more than *b*. Figure 3.2.2 shows the top 4 implicitly related concepts for the sample query.

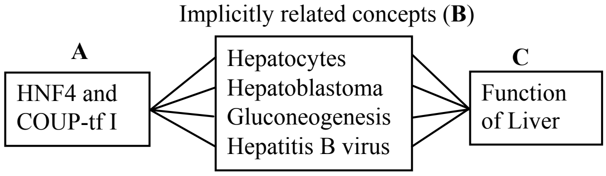

Figure 3.2.2 Top 4 implicitly related concepts for the query "How do interactions between HNF4 and COUP-TF1 suppress liver function?".

In Figure 3.2.2, the top 4 implicitly related concepts are all highly associated with "liver": "Hepatocytes" are liver cells; "Hepatoblastoma" is a malignant liver neoplasm occurring in young children; the vast majority of "Gluconeogenesis" takes place in the liver; and "Hepatitis B virus" is a virus that could cause serious damage to the function of liver.

The top-*k* ranked concepts in *B* are used for query expansion: if $I(b, A) \geq I(b, C)$, then *b* is considered as an implicit related concept of *A*. A document having *b* but not *A* will receive a partial weight of *A*. The expansion is similar for *C* when $I(b, A) < I(b, C)$.

3.3 Conceptual IR model

We now discuss our conceptual IR model. We first give the basic conceptual IR model in section 3.3.1. Then we explain how the domain-specific knowledge is incorporated in the model using query expansion in section 3.3.2. A pseudo-feedback strategy is introduced in section 3.3.3. In section 3.3.4, we give a strategy to improve the ranking by avoiding incorrect match of abbreviations.

3.3.1 Basic model

Given a query *q* and a document *d*, our model measures two similarities, concept similarity and word similarity:

$$sim(q,d) = (\underset{concept}{sim}(q,d),\ \underset{word}{sim}(q,d))$$

Concept similarity

Two vectors are derived from a query *q*,

$$q = (v_1, v_2)$$
$$v_1 = (c_{11}, c_{12}, ..., c_{1m})$$
$$v_2 = (c_{21}, c_{22}, ..., c_{2n})$$

where v_1 is a vector of concepts describing the biological object(s) and v_2 is a vector of concepts describing the biological process(es).

Given a vector of concepts *v*, let *s*(*v*) be the set of concepts in *v*. The weight of v_i is then measured by:

$$w(v_i) = \max\{\log\frac{N}{n_v} : s(v) \subseteq s(v_i) \text{ and } n_v > 0\}$$

where *v* is a vector that contains a subset of concepts in v_i and n_v is the number of documents having all the concepts in *v*.

The concept similarity between *q* and *d* is then computed by

$$\underset{concept}{sim}(q,d) = \sum_{i=1}^{2} \alpha_i \times w(v_i)$$

where α_i is a parameter to indicate the completeness of v_i that document *d* has covered. α_i is measured by:

$$\alpha_i = \frac{\sum_{c \in d \text{ and } c \in v_i} idf_c}{\sum_{c \in v_i} idf_c} \quad (3)$$

where idf_c is the inverse document frequency of concept *c*.

An example: suppose we have a query "How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity?". After identifying the concepts in the query, we have:

$$v_1 = (\text{'Nurr-77'})$$
$$v_2 = (\text{'T cells', 'spleen', 'autoimmunity', 'lymph nodes'})$$

Suppose that some document frequencies of different combinations of concepts are as follows:

25	df('Nurr-77')
0	df('T cells', 'spleen', 'autoimmunity', 'lymph nodes')
326	df('T cells', 'spleen', 'autoimmunity')
82	df('spleen', 'autoimmunity', 'lymph nodes')
147	df('T cells', 'autoimmunity', 'lymph nodes')
2332	df('T cells', 'spleen', 'lymph nodes')

The weight of v_i is then computed by (note that there does not exist a document having all the concepts in v_2):

$$w(v_1) = \log(N/25)$$
$$w(v_2) = \log(N/82)\ .$$

Now suppose a document *d* contains concepts 'Nurr-77', 'T cells', 'spleen', and 'lymph nodes', but not 'autoimmunity', then the value of parameter α_i is computed as follows:

$$\alpha_1 = 1$$
$$\alpha_2 = \frac{idf(\text{'T cells'}) + idf(\text{'spleen'}) + idf(\text{'lymph nodes'})}{idf(\text{'T cells'}) + idf(\text{'spleen'}) + idf(\text{'lymph nodes'}) + idf(\text{'autoimmunity'})}$$

Word similarity

The similarity between *q* and *d* on the word level is computed using Okapi [17]:

$$\underset{word}{sim}(q,d) = \sum_{w \in q} \log(\frac{N - n + 0.5}{n + 0.5}) \frac{(k_1 + 1)tf}{K + tf} \quad (4)$$

where *N* is the size of the document collection; *n* is the number of documents containing *w*; $K = k_1 \times ((1-b) + b \times dl/avdl)$ and $k_1 = 1.2$, *b*=0.75 are constants. *dl* is the document length of *d* and *avdl* is the average document length; *tf* is the term frequency of *w* within *d*.

The model

Given two documents d_1 and d_2, we say $sim(q,d_1) > sim(q,d_2)$ or d_1 will be ranked higher than d_2, with respect to the same query *q*, if either

1) $\underset{concept}{sim(q,d_1)} > \underset{concept}{sim(q,d_2)}$ **or**

2) $\underset{concept}{sim(q,d_1)} = \underset{concept}{sim(q,d_2)}$ and $\underset{word}{sim(q,d_1)} > \underset{word}{sim(q,d_2)}$

This conceptual IR model emphasizes the similarity on the concept level. A similar model but applied to non-biomedical domain has been given in [15].

3.3.2 Incorporating domain-specific knowledge

Given a concept *c*, a vector *u* is derived by incorporating its domain-specific knowledge:

$$u = (c, u_1, u_2, u_3)$$

where u_1 is a vector of its synonyms, hyponyms, and lexical variants; u_2 is a vector of its hypernyms; and u_3 is a vector of its implicitly related concepts. An occurrence of any term in u_1 will be counted as an occurrence of *c*. idf_c in Formula 3 is updated as:

$$idf_c = \log \frac{N}{|D_{c,u_1}|}$$

D_{c,u_1} is the set of documents having *c* or any term in u_1. The weight that a document *d* receives from *u* is given by:

$$\max\{w_t : t \in u \text{ and } t \in d\}$$

where $w_t = \beta \times idf_c$. The weighting factor β is an empirical tuning parameter determined as:

1. $\beta = 1$ if *t* is the original concept, its synonym, its hyponym, or its lexical variant;
2. $\beta = 0.95$ if *t* is a hypernym;
3. $\beta = 0.90 \times (k\text{–}i+1)/k$ if *t* is an implicitly related concept. *k* is the number of selected top ranked implicitly related concepts (see section 3.2.2); *i* is the position of *t* in the ranking of implicitly related concepts.

3.3.3 Pseudo-feedback

Pseudo-feedback is a technique commonly used to improve retrieval performance by adding new terms into the original query. We used a modified pseudo-feedback strategy described in [2].

Step 1. Let *C* be the set of concepts in the top 15 ranked documents. For each concept *c* in *C*, compute the similarity between *c* and the query *q*, the computation of *sim*(*q,c*) can be found in [2].

Step 2. The top-*k* ranked concepts by *sim*(*q,c*) are selected.

Step 3. Associate each selected concept *c'* with the concept c_q in *q* that 1) has the same semantic type as *c'*, and 2) is most related to *c'* among all the concepts in *q*. The association between *c'* and c_q is computed by:

$$I(c', c_q) = \log \frac{\mathrm{P}(c', c_q)}{\mathrm{P}(c')\mathrm{P}(c_q)}$$

where P(*x*) = *n*/*N*, *n* is the number of documents having *x* and *N* is the size of the document collection. A document having *c'* but not c_q receives a weight given by: $(0.5 \times (k\text{-}i+1)/k) \times idf_{c_q}$, where *i* is the position of *c'* in the ranking of step 2.

3.3.4 Avoid incorrect match of abbreviations

Some gene symbols are very short and thus ambiguous. For example, the gene symbol "APC" could be the abbreviation for many non-gene long-forms, such as "air pollution control", "aerobic plate count", or "argon plasma coagulation". This step is to avoid incorrect match of abbreviations in the top ranked documents.

Given an abbreviation *X* with the long-form *L* in the query, we scan the top-*k* ranked (*k*=1000) documents and when a document is found with *X*, we compare *L* with all the long-forms of *X* in that document. If none of these long-forms is equal or close to *L* (i.e., the edit distance between *L* and the long-form of *X* in that document is 1 or 2), then the concept similarity of *X* is subtracted.

3.4 Passage extraction

The goal of passage extraction is to highlight the most relevant fragments of text in paragraphs. A passage is defined as any span of text that does not include the HTML paragraph tag (i.e., <P> or </P>). A passage could be a part of a sentence, a sentence, a set of consecutive sentences or a paragraph (i.e., the whole span of text that are inside of <P> and </P> HTML tags). It is also possible to have more than one relevant passage in a single paragraph. Our strategy for passage extraction assumes that the optimal passage(s) in a paragraph should have all the query concepts that the whole paragraph has. Also they should have higher density of query concepts than other fragments of text in the paragraph.

Suppose we have a query *q* and a paragraph *p* represented by a sequence of sentences $p = s_1 s_2 \ldots s_n$. Let *C* be the set of concepts in *q* that occur in *p* and $S = \Phi$.

Step 1. For each sequence of consecutive sentences $s_i s_{i+1} \ldots s_j$, $1 \le i \le j \le n$, let $S = S \cup \{s_i s_{i+1} \ldots s_j\}$ if $s_i s_{i+1} \ldots s_j$ satisfies that:

1) Every query concept in *C* occurs in $s_i s_{i+1} \ldots s_j$ and

2) There does not exist *k*, such that $i < k < j$ and every query concept in *C* occurs in $s_i s_{i+1} \ldots s_k$ or $s_{k+1} s_{k+2} \ldots s_j$.

Condition 1 requires $s_i s_{i+1} \ldots s_j$ having all the query concepts in *p* and condition 2 requires $s_i s_{i+1} \ldots s_j$ be the minimal.

Step 2. Let $L = \min\{j - i + 1 : s_i s_{i+1} \ldots s_j \in S\}$. For every $s_i s_{i+1} \ldots s_j$ in *S*, let $S = S - \{s_i s_{i+1} \ldots s_j\}$ if $(j - i + 1) > L$. This step is to remove those sequences of sentences in *S* that have lower density of query concepts.

Step 3. For every two sequences of consecutive sentences $s_{i_1} s_{i_1+1} \ldots s_{j_1} \in S$, and $s_{i_2} s_{i_2+1} \ldots s_{j_2} \in S$, if

$$\begin{aligned} & i_1 \le i_2, j_1 \le j_2 \quad \text{and} \\ & i_2 \le j_1 + 1 \end{aligned} \tag{5}$$

then do

$$\begin{aligned} S &= S \cup \{s_{i_1} s_{i_1+1} \ldots s_{j_2}\} \\ S &= S - \{s_{i_1} s_{i_1+1} \ldots s_{j1}\} \\ S &= S - \{s_{i_2} s_{i_2+1} \ldots s_{j_2}\} \end{aligned}$$

Repeat this step until for every two sequences of consecutive sentences in *S*, condition (5) does not apply. This step is to merge those sequences of sentences in *S* that are adjacent or overlapped.

Finally the remaining sequences of sentences in *S* are returned as the optimal passages in the paragraph *p* with respect to the query.

4. EXPERIMENTAL RESULTS

The evaluation of our techniques and the experimental results are given in this section. We first describe the datasets and evaluation metrics used in our experiments and then present the results.

4.1 Data sets and evaluation metrics

Our experiments were performed on the platform of the Genomics track of TREC 2006. The document collection contains 162,259 full-text documents from 49 Highwire biomedical journals. The set of queries consists of 28 queries collected from real biologists.

The performance is measured on three different levels (passage, aspect, and document) to provide better insight on how the question is answered from different perspectives. **Passage MAP:** As described in [8], this is a character-based precision calculated as follows: "At each relevant retrieved passage, precision will be computed as the fraction of characters overlapping with the gold standard passages divided by the total number of characters included in all nominated passages from this system for the topic up until that point. Similar to regular MAP, relevant passages that were not retrieved will be added into the calculation as well, with precision set to 0 for relevant passages not retrieved. Then the mean of these average precisions over all topics will be calculated to compute the mean average passage precision". **Aspect MAP**: A question could be addressed from different aspects. For example, the question "what is the role of gene PRNP in the Mad cow disease?" could be answered from aspects like "Diagnosis", "Neurologic manifestations", or "Prions/Genetics". This measure indicates how comprehensive the question is answered. **Document MAP**: This is the standard IR measure. The precision is measured at every point where a relevant document is obtained and then averaged over all relevant documents to obtain the average precision for a given query. For a set of queries, the mean of the average precision for all queries is the MAP of that IR system.

The output of the system is a list of passages ranked according to their similarities with the query. The performances on the three levels are then calculated based on the ranking of the passages.

4.2 Results

The Wilcoxon signed-rank test was employed to determine the statistical significance of the results. In the tables of the following sections, statistically significant improvements (at the 5% level) are marked with an asterisk.

4.2.1 Conceptual IR model vs. term-based model

The initial baseline was established using word similarity only computed by the Okapi (Formula 4). Another run based on our basic conceptual IR model was performed without using query expansion, pseudo-feedback, or abbreviation correction. The experimental result is shown in Table 4.2.1. Our basic conceptual IR model significantly outperforms the Okapi on all three levels, which suggests that, although it requires additional efforts to identify concepts, retrieval on the concept level can achieve substantial improvements over purely term-based retrieval model.

4.2.2 Contribution of different types of knowledge

A series of experiments were performed to examine how each type of domain-specific knowledge contributes to the retrieval performance. A new baseline was established using the basic conceptual IR model without incorporating any type of domain-specific knowledge. Then five runs were conducted by adding each individual type of domain-specific knowledge. We also conducted a run by adding all types of domain-specific knowledge. Results of these experiments are shown in Table 4.2.2.

We found that any available type of domain-specific knowledge improved the performance in passage retrieval. The biggest improvement comes from the lexical variants, which is consistent with the result reported in [3]. This result also indicates that biologists are likely to use different variants of the same concept according to their own writing preferences and these variants might not be collected in the existing biomedical thesauruses. It also suggests that the biomedical IR systems can benefit from the domain-specific knowledge extracted from the literature by text mining systems.

Synonyms provided the second biggest improvement. Hypernyms, hyponyms, and implicitly related concepts provided similar degrees of improvement. The overall performance is an accumulative result of adding different types of domain-specific knowledge and it is better than any individual addition. It is clearly shown that the performance is significantly improved (107% on passage level, 63.1% on aspect level, and 49.6% on document level) when the domain-specific knowledge is appropriately incorporated. Although it is not explicitly shown in Table 4.2.3, different types of domain-specific knowledge affect different subsets of queries. More specifically, each of these types (with the exception of "the lexical variants" which affects a large number of queries) affects only a few queries. But for those affected queries, their improvement is significant. As a consequence, the accumulative improvement is very significant.

4.2.3 Pseudo-feedback and abbreviation correction

Using the "Baseline+All" in Table 4.2.2 as a new baseline, the contribution of abbreviation correction and pseudo-feedback is given in Table 4.2.3. There is little improvement by avoiding incorrect matching of abbreviations. The pseudo-feedback contributed about 4.6% improvement in passage retrieval.

4.2.4 Performance compared with best-reported results

We compared our result with the results reported in the Genomics track of TREC 2006 [8] on the conditions that 1) systems are automatic systems and 2) passages are extracted from paragraphs. The performance of our system relative to the best reported results is shown in Table 4.2.4 (in TREC 2006, some systems returned the whole paragraphs as passages. As a consequence, excellent retrieval results were obtained on document and aspect levels at the expense of performance on the passage level. We do not include the results of such systems here).

Table 4.2.4 Performance compared with best-reported results.

	Passage MAP	Aspect MAP	Document MAP
Best reported results	0.1486	0.3492	0.5320
Our results	0.1823	0.3811	0.5391
Improvement	22.68%	9.14%	1.33%

The best reported results in the first row of Table 4.2.4 on three levels (passage, aspect, and document) are from different systems. Our result is from a single run on passage retrieval in which it is better than the best reported result by 22.68% in passage retrieval and at the same time, 9.14% better in aspect retrieval, and 1.33% better in document retrieval (Since the average precision of each individual query was not reported, we can not apply the Wilcoxon signed-rank test to calculate the significance of difference between our performance and the best reported result.).

Table 4.2.1 Basic conceptual IR model vs. term-based model

Run	Passage		Aspect		Document	
	MAP	Imprvd qs # (%)	MAP	Imprvd qs # (%)	MAP	Imprvd qs # (%)
Okapi	0.064	N/A	0.175	N/A	0.285	N/A
Basic conceptual IR model	0.084* (+31.3%)	17 (65.4%)	0.233* (+33.1%)	12 (46.2%)	0.359* (+26.0%)	15 (57.7%)

Table 4.2.2 Contribution of different types of domain-specific knowledge

Run	Passage		Aspect		Document	
	MAP	Imprvd qs # (%)	MAP	Imprvd qs # (%)	MAP	Imprvd qs # (%)
Baseline = Basic conceptual IR model	0.084	N/A	0.233	N/A	0.359	N/A
Baseline+Synonyms	0.105 (+25%)	11 (42.3%)	0.246 (+5.6%)	9 (34.6%)	0.420 (+17%)	13 (50%)
Baseline+Hypernyms	0.088 (+4.8%)	11 (42.3%)	0.225 (-3.4%)	9 (34.6%)	0.390 (+8.6%)	16 (61.5%)
Baseline+Hyponyms	0.087 (+3.6%)	10 (38.5%)	0.217 (-6.9%)	7 (26.9%)	0.389 (+8.4%)	10 (38.5%)
Baseline+Variants	0.150* (+78.6%)	16 (61.5%)	0.348* (+49.4%)	13 (50%)	0.495* (+37.9%)	10 (38.5%)
Baseline+Related	0.086 (+2.4%)	9 (34.6%)	0.220 (-5.6%)	9 (34.6%)	0.387 (+7.8%)	13 (50%)
Baseline+All	**0.174* (107%)**	25 (96.2%)	**0.380* (+63.1%)**	19 (73.1%)	**0.537* (+49.6%)**	14 (53.8%)

Table 4.2.3 Contribution of abbreviation correction and pseudo-feedback

Run	Passage		Aspect		Document	
	MAP	Imprvd qs # (%)	MAP	Imprvd qs # (%)	MAP	Imprvd qs # (%)
Baseline+All	0.174	N/A	0.380	N/A	0.537	N/A
Baseline+All+Abbr	0.175 (+0.6%)	5 (19.2%)	0.375 (-1.3%)	4 (15.4%)	0.535 (-0.4%)	4 (15.4%)
Baseline+All+Abbr+PF	0.182 (+4.6%)	10 (38.5%)	0.381 (+0.3%)	6 (23.1%)	0.539 (+0.4%)	9 (34.6%)

A separate experiment has been done using a second testbed, the Ad Hoc Task of TREC Genomics 2005, to evaluate our knowledge-intensive conceptual IR model for document retrieval of biomedical literature. The overall performance in terms of MAP is 35.50%, which is about 22.92% above the best reported result [9]. Notice that the performance was only measured on the document level for the Ad Hoc Task of TREC Genomics 2005.

5. RELATED WORKS

Many studies used manually-crafted thesauruses or knowledge databases created by text mining systems to improve retrieval effectiveness based on either word-statistical retrieval systems or conceptual retrieval systems.

[11][1] assessed query expansion using the UMLS Metathesaurus. Based on a word-statistical retrieval system, [11] used definitions and different types of thesaurus relationships for query expansion and a deteriorated performance was reported. [1] expanded queries with phrases and UMLS concepts determined by the MetaMap, a program which maps biomedical text to UMLS concepts, and no significant improvement was shown. We used MeSH, Entrez gene, and other non-thesaurus knowledge resources such as an abbreviation database for query expansion. A critical difference between our work and those in [11][1] is that our retrieval model is based on concepts, not on individual words.

The Genomics track in TREC provides a common platform to evaluate methods and techniques proposed by various groups for biomedical information retrieval. As summarized in [8][9][10], many groups utilized domain-specific knowledge to improve retrieval effectiveness. Among these groups, [3] assessed both thesaurus-based knowledge, such as gene information, and non thesaurus-based knowledge, such as lexical variants of gene symbols, for query expansion. They have shown that query expansion with acronyms and lexical variants of gene symbols produced the biggest improvement, whereas, the query expansion with gene information from gene databases deteriorated the performance. [21] used a similar approach for generating lexical variants of gene symbols and reported significant improvements. Our system utilized more types of domain-specific knowledge, including hyponyms, hypernyms and implicitly related concepts. In addition, under the conceptual retrieval framework, we examined more comprehensively the effects of different types of domain-specific knowledge in performance contribution.

[20][15] utilized WordNet, a database of English words and their lexical relationships developed by Princeton University, for query expansion in the non-biomedical domain. In their studies, queries were expanded using the lexical semantic relations such as synonyms, hypernyms, or hyponyms. Little benefit has been shown in [20]. This has been due to ambiguity of the query terms which have different meanings in different contexts. When these synonyms having multiple meanings are added to the query, substantial irrelevant documents are retrieved. In the biomedical domain, this kind of ambiguity of query terms is relatively less frequent, because, although the abbreviations are highly ambiguous, general biomedical concepts usually have only one meaning in the thesaurus, such as UMLS, whereas a term in WordNet usually have multiple meanings (represented as synsets in WordNet). Besides, we have implemented a post-ranking step to reduce the number of incorrect matches of abbreviations, which will hopefully decrease the negative impact caused by the abbreviation ambiguity. Besides, we have implemented a post-ranking step to reduce the number of incorrect matches of abbreviations, which will hopefully decrease the negative impact caused by the abbreviation ambiguity. The retrieval model in [15] emphasized the similarity between a query and a document on the phrase level assuming that phrases are more important than individual words when retrieving documents. Although the assumption is similar, our conceptual model is based on the biomedical concepts, not phrases.

[13] presented a good study of the role of knowledge in the document retrieval of clinical medicine. They have shown that appropriate use of semantic knowledge in a conceptual retrieval framework can yield substantial improvements. Although the retrieval model is similar, we made a study in the domain of genomics, in which the problem structure and task knowledge is not as well-defined as in the domain of clinical medicine [18]. Also, our similarity function is very different from that in [13].

In summary, our approach differs from previous works in four important ways: First, we present a case study of conceptual retrieval in the domain of genomics, where many knowledge resources can be used to improve the performance of biomedical IR systems. Second, we have studied more types of domain-specific knowledge than previous researchers and carried out more comprehensive experiments to look into the effects of different types of domain-specific knowledge in performance contribution. Third, although some of the techniques seem similar to previously published ones, they are actually quite different in details. For example, in our pseudo-feedback process, we require that the unit of feedback is a concept and the concept has to be of the same semantic type as a query concept. This is to ensure that our conceptual model of retrieval can be applied. As another example, the way in which implicitly related concepts are extracted in this paper is significantly different from that given in [19]. Finally, our conceptual IR model is actually based on complex concepts because some biomedical meanings, such as biological processes, are represented by multiple simple concepts.

6. CONCLUSION

This paper proposed a conceptual approach to utilize domain-specific knowledge in an IR system to improve its effectiveness in retrieving biomedical literature. We specified five different types of domain-specific knowledge (i.e., synonyms, hyponyms, hypernyms, lexical variants, and implicitly related concepts) and examined their effects in performance contribution. We also evaluated other two techniques, pseudo-feedback and abbreviation correction. Experimental results have shown that appropriate use of domain-specific knowledge in a conceptual IR model yields significant improvements (23%) in passage retrieval over the best known results. In our future work, we will explore the use of other existing knowledge resources, such as UMLS and the Wikipedia, and evaluate techniques such as disambiguation of gene symbols for improving retrieval effectiveness. The application of our conceptual IR model in other domains such as clinical medicine will be investigated.

7. ACKNOWLEDGMENTS

We thank the anonymous SIGIR 2007 reviewers for their useful comments. We would like to thank Dr. David Featherstone for his insightful discussion.

8. REFERENCES

[1] Aronson A.R., Rindflesch T.C. Query expansion using the UMLS Metathesaurus. *Proc AMIA Annu Fall Symp*. 1997. 485-9.

[2] Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. Addison-Wesley, 1999, 129-131.

[3] Buttcher S., Clarke C.L.A., Cormack G.V. Domain-specific synonym expansion and validation for biomedical information retrieval (MultiText experiments for TREC 2004). *TREC'04*.

[4] Chang J.T., Schutze H., Altman R.B. Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*. 2002 9(6).

[5] Church K.W., Hanks P. Word association norms, mutual information and lexicography. *Computational Linguistics*. 1990;16:22, C29.

[6] Fontelo P., Liu F., Ackerman M. askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. *BMC Med Inform Decis Mak*. 2005 Mar 10;5(1):5.

[7] Fukuda K., Tamura A., Tsunoda T., Takagi T. Toward information extraction: identifying protein names from biological papers. *Pac Symp Biocomput*. 1998;:707-18.

[8] Hersh W.R., and etc. TREC 2006 Genomics Track Overview. *TREC'06*.

[9] Hersh W.R., and etc. TREC 2005 Genomics Track Overview. In *TREC'05*.

[10] Hersh W.R., and etc. TREC 2004 Genomics Track Overview. In *TREC'04*.

[11] Hersh W.R., Price S., Donohoe L. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. *Proc AMIA Symp*. 344-8. 2000.

[12] Levenshtein, V. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics* - Doklady 10, 10 (1996), 707-710.

[13] Lin J., Demner-Fushman D. The Role of Knowledge in Conceptual Retrieval: A Study in the Domain of Clinical Medicine. *SIGIR'06*. 99-06.

[14] Lindberg D., Humphreys B., and McCray A. The Unified Medical Language System. *Methods of Information in Medicine*. 32(4):281-291, 1993.

[15] Liu S., Liu F., Yu C., and Meng W.Y. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases. *SIGIR'04*. 266-272

[16] Proux D., Rechenmann F., Julliard L., Pillet V.V., Jacq B. Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Inform Ser Workshop Genome Inform*. 1998;9:72-80.

[17] Robertson S.E., Walker S. Okapi/Keenbow at TREC-8. *NIST Special Publication 500-246: TREC 8.*

[18] Sackett D.L., and etc. Evidence-Based Medicine: How to Practice and Teach EBM. *Churchill Livingstone*. Second edition, 2000.

[19] Swanson,D.R., Smalheiser,N.R. An interactive system for finding complemen-tary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 1997; 91,183-203.

[20] Voorhees E. Query expansion using lexical-semantic relations. *SIGIR* 1994. 61-9

[21] Zhong M., Huang X.J. Concept-based biomedical text retrieval. *SIGIR'06*. 723-4

[22] Zhou W., Torvik V.I., Smalheiser N.R. ADAM: Another Database of Abbreviations in MEDLINE. *Bioinformatics*. 2006; 22(22): 2813-2818.