Experiment report of TREC 2005 Genomics track ad hoc retrieval task

Wei Zhou, Clement Yu

Database and Information System Lab
Computer Science Department, University of Illinois at Chicago
{wzhou1,yu}cs.uic.edu

Abstract

This report describes the experiments we have conducted on the ad hoc retrieval task of Genomics track at TREC 2005. In the experiment, a number of different techniques were employed, including Porter stemming, MeSH term and gene name identification, Okapi, weighting schemes, query expansion, and concept-based ranking strategy. The results on sample topics are reported. Future improvements, such as utilizing domain-specific knowledge, gene name disambiguation, and pseudo-feedback are discussed.

Introduction

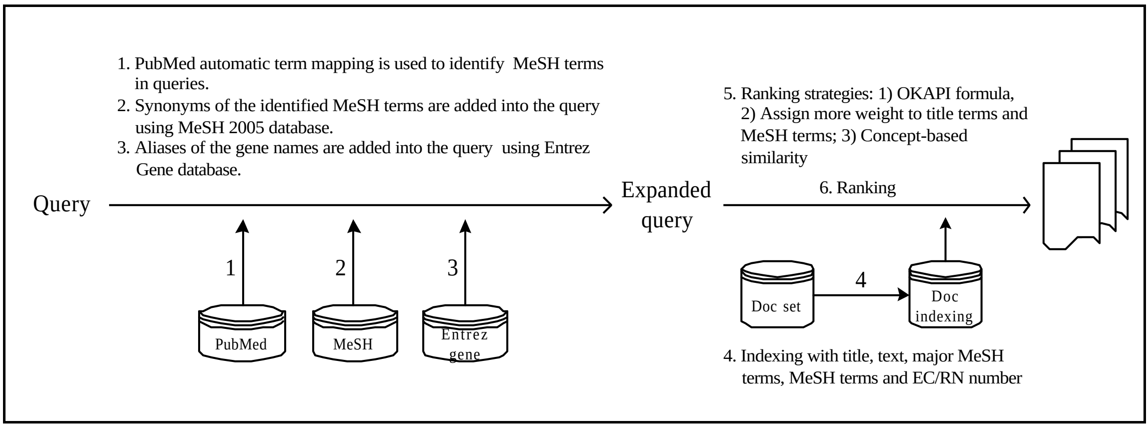
The purpose of the ad hoc retrieval task of Genomics track at TREC 2005 is to provide systems with better-defined queries (each query is derived from a generic topic template which includes a certain number of semantic types) for finding genomics information [1]. The topics in the 2005 ad hoc retrieval task were collected from real biologists and are more structured than the mostly free-form topics from the 2004 track. Totally 50 topics, derived from 5 generic topic templates (GTTs)[1], each of which has 10 instances, are provided. Participants are expected to build systems to find information about these 50 topics from a 10-year MEDLINE subset, which includes 4,591,008 documents from year 1994 to 2003.

This report will first introduce our retrieval strategy, including preprocessing, MeSH term identification, OKAPI, query expansion, and concept-based ranking. Next, the experiment results on the sample queries are reported. In the section of conclusion and future work, some improvements are discussed.

[1] A sample GTT is: " Find articles describing the role of a gene involved in a given disease." The semantic types in the GTT are underlined.

Retrieval strategy

Figure 1 shows the whole query processing procedure. In the process, PubMed, MeSH 2005, and Entrez gene are used to expand the query.


1. PubMed automatic term mapping is used to identify MeSH terms in queries.
2. Synonyms of the identified MeSH terms are added into the query using MeSH 2005 database.
3. Aliases of the gene names are added into the query using Entrez Gene database.
5. Ranking strategies: 1) OKAPI formula, 2) Assign more weight to title terms and MeSH terms; 3) Concept-based similarity
Query
Expanded query
6. Ranking
1
2
3
PubMed
MeSH
Entrez gene
4
Doc set
Doc indexing
4. Indexing with title, text, major MeSH terms, MeSH terms and EC/RN number

F
igure 1: query processing procedure

1. Preprocessing

Both the documents and queries were stoplisted and stemmed. Uninformative terms in the documents and queries were removed according to the PubMed official stop words list, which includes 137 most frequently used terms in PubMed. The Porter stemmer was used to handle lexical variants. For each document or a MEDLINE record, the following fields were used to build a word-level inverted document indexing: Title, Abstract, EC/RN Number[2], and MeSH[3]. On the other hand, it has been observed that title field and MeSH field are likely to contain those terms that describe the document's topics, especially the Major MeSH terms[4] [5]. Based on this observation, a concept-level indexing was created on title, Major MeSH terms, EC/RN Number, and MeSH terms. Concepts defined in our experiment are MeSH terms, their synonyms, and EC/RN numbers. The format of the concept-level indexing is as follows:

ConceptID$_i$, FieldType, DocFrq, PubMedID$_1$| PubMedID$_2$|……| PubMedID$_k$

[2] EC/RN number: Number assigned by the Enzyme Commission to designate a particular enzyme or by the Chemical Abstracts Service for Registry Numbers.

[3] MeSH: a controlled vocabulary of biomedical terms that is used to describe the subject of each journal article in MEDLINE

[4] Major MeSH term: A MeSH term that is one of the main topics discussed in the article denoted by an asterisk on the MeSH term or MeSH/Subheading combination, e.g., Cytokines/physiology*

where ConceptID_i is the ID number of each concept. FieldType indicates where the concept occurs. It can be *TI* (Title), *MJR* (Major MeSH term), *ECRN* (EC/RN Number), OR *MH* (MeSH term). DocFrq is the total number of documents having the concept. PubMedID_1| PubMedID_2|……| PubMedID_k is the list of those documents.

2. MeSH term Identification

The PubMed Automatic Term Mapping [4] is used to retrieve MeSH terms. The mapping between terms in user's queries and MeSH terms in PubMed is done through its *MeSH Translation Table*. It compares terms from a query with lists of terms in its translation table, which contains:

1) MeSH Headings.
2) MeSH Subheadings: Used by NLM to further describe a certain aspect of a MeSH heading. For example, the MeSH heading "Liver" may be qualified with the subheading "drug-effects" (the PubMed query will be "liver [mh] AND de[sh]") to indicate that the article is not about the liver in general, but about the effect of drugs on the liver.
3) Synonyms of MeSH headings and subheadings in UMLS.

In our experiment, the queries were searched in PubMed through its *ESearch* E-utility and MeSH terms were extracted by parsing the returned XML file. For example, Table 1 shows the PubMed translation of the query "purification of rat IgM".

Table 1: An example of PubMed automatic term mapping (three MeSH terms are extracted: "immunoglobulin m", "isolation and purification", and "rats". These MeSH terms will be used for query expansion.)

Term	**PubMed translation**
IgM	("immunoglobulin m"[TIAB] NOT Medline[SB]) OR "immunoglobulin m"[MeSH Terms] OR IgM[Text Word]
Purification	"isolation and purification"[Subheading] OR Purification[Text Word]
rat	("rats"[TIAB] NOT Medline[SB]) OR "rats"[MeSH Terms] OR rat[Text Word]

TIAB: Words and numbers included in the title, abstract, and other abstract of a citation.

NOT Medline[SB]: PubMed will retrieve additional citations that have not been indexed for MEDLINE, e.g., in-process and OLDMEDLINE.

Text Word: Includes all words and numbers in the title, abstract, other abstract, MeSH terms, MeSH Subheadings, chemical substance names, personal name as subject, MEDLINE Secondary Source, and Other Terms typically non-MeSH subject terms (keywords), including NASA Space Flight Mission, assigned by an organization other than NLM.

3. Okapi weighting

Okapi relevance scoring formula [6] is known to embody a good model of relevance based upon term occurrences within text documents. The version in our experiment is as follows:

$$\sum_{T \in Q} w_T \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \qquad \textbf{(1)}$$

where Q is a query, containing term T

w_T is the weight of T in Q

$$w_T = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \qquad \textbf{(2)}$$

N=4,591,008 is the number of documents in the collection.
n is the number of documents containing the term.
R is the number of documents known to be relevant to a specific topic and *r* is the number of relevant documents containing the term. Since R and r are not known in our experiment, formula (2) can be simplified as:

$$w_T = \log(\frac{N-n+0.5}{n+0.5}) \qquad \textbf{(3)}$$

K is $k_1 \times ((1-b)+b \times \frac{dl}{avdl})$
k_1=1.2, b=0.75 are the constants used in experiments reported here. *dl* is the length of the document and *avdl*=183.5554 is the average document length.
tf is the frequency of occurrence of the term within a specific document.
qtf is the frequency of the term within the topic from which *Q* was derived. *qtf* =1 is a constant in our experiment. So formula (1) can be modified as:

$$\sum_{T \in Q} w_T \frac{(k_1+1)tf}{K+tf} \qquad \textbf{(4)}$$

For each term, its OKAPI weight in each document ($w_T \frac{(k_1+1)tf}{K+tf}$) was pre-computed and stored. The data structure is as follows:

Term, DocFrq, PubMedID$_1$:OKAPIscore| PubMedID$_2$:OKAPIscore |……| PubMedID$_k$:OKAPIscore

4. Query expansion

Once the MeSH terms in the original queries were extracted, the MeSH 2005 database was used to expand the original query. The synonyms (also known as entry terms) for MeSH terms were added into the query. For example:

MeSH term: Immunoglobulin M
Added Synonyms: “Gamma Globulin, 19S”, IgM, IgM1, IgM2

As long as any of these terms appears in a document, we consider this concept appears in that document. Besides the MeSH terms, the gene names were also expanded by adding their aliases in the Entrez Gene database[5] [3]. For example:

Gene name: BRCA1
Added aliases: BRCAI, IRIS, PSCP, RNF53

Like the MeSH terms, occurring of any of these terms in a document will be counted as one occurrence of this gene in that document.

5. Ranking

The similarity between a document and a query has two components: one is concept similarity and the other term similarity:

$$sim(Q, D) = (concept\_sim, term\_sim)$$

The concept similarity is the sum of all the concept weights in the query.

$$concept\_sim = \sum_{C \in Q} w_C$$

where Q is a query, containing concepts C

w_C is the weight of C in Q

$$w_C = \log \frac{N}{n}$$

where N=4,591,008 is the number of documents in the collection. n is the number of documents containing the concept or any of its synonyms. For example, gene *BRCA1* is a concept and it has a set of synonyms (*BRCAI*, *IRIS*, *PSCP*, *RNF53*). The number of documents having concept *BRCA1* is computed with the following PubMed Boolean search:

BRCA1[Text Word] OR BRCAI[Text Word] OR IRIS[Text Word] OR PSCP[Text Word] OR RNF53[Text Word].

Totally 2,267 documents are retrieved. This number is the document frequency of concept *BRCA1*. Notice that *BRCAI*, *IRIS*, *PSCP*, *RNF53* are assigned the same weight as *BRCA1*. For example, document D_1 has *BRCA1*, D_2 has *IRIS*, and the query has *BRCA1*. According to our weighting strategy, D_1 and D_2 get the same weight on this concept.

Notice that concepts in the title, MeSH, or EC/RN field are likely to describe the topics of a

[5] Entrez Gene supercedes LocusLink as the most prominent source of publicly available information on genes. It provides detailed information about the function and position of genes. Gene aliases are unofficial symbols and descriptions that have been used for this gene and its products.

document. Based on this observation, more weight is assigned to those concepts occurring in these fields. For example, suppose concept C is an indexing MeSH term in document D, the weight of C in D will be

$$w_C = (1 + a) \times \log(\frac{N}{n})$$

a is a parameter to adjust. In our experiment, a is equal to 0.3 when C is a major MeSH term and 0.2 when C is a concept in the title, MeSH, or EC/RN number field.

$term\_sim$ is computed with formula (4) ($\sum_{T \in Q} w_T \frac{(k_1 + 1)tf}{K + tf}$).

Ranking is concept-based. Consider, for a query Q, document D_1 and D_2, having similarities (x_1, y_1) and (x_2, y_2), respectively, D_1 will be ranked higher than D_2 if either (1) $x_1 > x_2$, or (2) $x_1 = x_2 ; y_1 > y_2$

Experiment result

10 sample topics, with two coming from each genetic topic template, were provided for participants to evaluate their systems. We use the *trec_eval* program to compute the 11-point average precision. Table 2 shows that our method has achieved 0.4002 average precision.

Analysis:

Query 91: “Describe the procedure or methods for GST fusion protein expression in Sf9 insect cells.”
43/56 relevant documents were missed. These 43 documents have "baculoviridae" or “baculovirus”, instead of "Sf9".

Query 93: “Provide information about the role of the gene DRD4 in the disease Alcoholism.”
4/37 relevant documents were missed. 1/4 (9691193) has “d4 dopamine receptor” , instead of “receptors, dopamine d4”. 1/4 (9285967) has ‘dopamine D4 receptor’, instead of ‘receptors, dopamine d4”. 2/4 (9406938, 11697748) have “dopamine receptor” and “d4”, but they are not adjacent.

Query 94: “Provide information on the role of the gene HMG in the process of chromatin restructuring and transcriptional regulation.”
40/66 relevant documents were missed. These 42 documents have “transcriptional activation” or “transcriptional activity”, or “transcriptional activator”, instead of “transcriptional regulation”.

Query 95: "Provide information on the role of the gene Insulin receptor gene in the process of signaling tumorigenesis."
8/30 relevant documents were missed. 1/8 (14556818) doesn't have "signaling tumorigenesis". For the other 7 documents, both "Insulin receptor" and "signaling tumorigenesis" can be found. However, they either only occur in the abstracts, or in the MeSH field (not as major MeSH terms).

Query 98: "Provide information about Mutation of Ret in thyroid function."
12/220 relevant documents were missed. 9/12 documents have "Ret oncogene protein, Drosophila" as an EC/RN number, but not " Ret" . (You get 1,152 documents when you search "Ret oncogene protein, Drosophila"[Substance Name] in PubMed. However, no documents are returned when you search "Ret" [Substance Name].) 3/12 documents don't have concept "mutation".

Table 2: Experiment result on the sample topics. (Not all the relevant documents for each query are retrieved; DR documents are used as the *qrel* file to compute the average precision.)

Query id	# unretrieved relevant documents	# relevant documents	11-point Avg precision
90	0	28	0.3358
91	43	56	0.1738
92	0	3	0.8409
93	4	37	0.7298
94	40	66	0.1977
95	8	30	0.0910
96	0	4	0.5527
97	0	5	0.1244
98	12	220	0.3346
99	0	64	0.6208
		Avg:	0.4002

For the official topics, only those documents that have at least one of the concepts in the query were retrieved For each topic, up to 1,000 documents were retrieved. Table 3 gives the results:

Table 3: Results of official queries

Summary Statistics	
Run ID	UICgen1
Run Type	automatic
Number of Topics	49

Total number of documents over all topics	
Retrieved	46918
Relevant	4584
Rel-ret	2839

Recall Level Averages	
Recall	Precision
0.00	0.531
0.10	0.3622
0.20	0.2924
0.30	0.2334
0.40	0.1877
0.50	0.1486
0.60	0.1173
0.70	0.0965
0.80	0.0877
0.90	0.0317
1.00	0.0059

Document Leval Averages	
	Precision
At 5 docs	0.3184
At 10 docs	0.3082
At 15 docs	0.3007
At 20 docs	0.2898
At 30 docs	0.2667
At 100 docs	0.1941
At 200 docs	0.1373
At 500 docs	0.0913
At 1000 docs	0.0579

R-Precision	0.2079
bpref	0.5474

Mean Average Precision	
non-interpolated	0.1738

Conclusions and future work

1. Domain-specific knowledge

The major difficulty comes from the lack of domain-specific knowledge. Each query is asking for information about a certain aspect of a topic. Table 4 is the relevance judgment for each topic template in the TREC 2005 Genomics Track protocol:

The retrieved document must describe a specific role of gene, a specific gene interaction, or etc., not the concepts generally. For example, query 111 is looking for information about the role of the gene *PRNP* in the disease *Mad Cow Disease*. Those documents that have mentioned *PRNP* and *Mad Cow Disease* may talk about the clinic trial, not the role of gene *PRNP*. We propose two ways of handling this problem.

Table 4: Relevance judgment of the sample topics

ID	Topic Template	Relevance judgment
1	Find articles describing standard methods or protocols for doing some sort of experiment or procedure.	Relevant article must describe how to conduct, adjust, or improve a standard, a, new method, or a protocol for doing some sort of experiment or procedure.
2	Find articles describing the role of a gene involved in a given disease.	Relevant article must describe some specific role of the gene in the stated disease.
3	Find articles describing the role of a gene in a specific biological process.	Relevant article must describe some specific role of the gene in the stated biological process.
4	Find articles describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease.	Relevant article must describe a specific interaction (e.g., promote, suppress, inhibit, etc.) between two or more genes in the stated function of the organ or the disease.
5	Find articles describing one or more mutations of a given gene and its biological impact.	Relevant article must describe a mutation of the stated gene and the particular biological impact(s) that the mutation has been found to have.

1) Using MeSH subheadings

MeSH subheadings are used to further describe a certain aspect of a MeSH heading. For example, the MeSH heading "Liver" may be qualified with the subheading "drug-effects" to indicate tha t the article is not about the liver in general, but about the effect of drugs on the liver.

For the query, using MeSH subheadings will help identify a specific aspect of the query topic. For template 1, the appropriate subheading is "Methods", which is used for MEDLINE indexing "with techniques, procedures, and programs for methods" [2]. For other templates (2, 3, 4, and 5), the suitable subheading is "Genetics", which is used "for mechanisms of heredity and the genetics of organisms, for the genetic basis of normal and pathologic states, and for the genetic aspects of endogenous chemicals. It includes biochemical and molecular influence on genetic material."[2]. Table 5 lists the combined MeSH heading/MeSH subheading for sample query 92-99:

To do it automatically, we first identify MeSH terms in the query (see section "MeSH term Identification"). If the MeSH term can be qualified with subheading "Genetics" (Notice: each MeSH term is associated with a list of subheadings), we assign more weight to those documents that are indexed with the combination of that MeSH term and "Genetics"

Table 5: MeSH subheadings for sample topics

ID	Topic Template
92	Ribosomal Proteins/genetics Proto-Oncogen Proteins/genetics
93	Alcoholism/genetics Alcohol Drinking/genetics Receptors, Dopamine D2/genetics
94	Hydroxymethylglutary1 CoA Reductases/genetics
95	Receptor, Insulin/genetics
96	N/A
97	N/A
98	Receptor Protein-Tyrosine Kinases/genetics
99	Methyltransferases/genetics

2) Using UMLS semantic relations

Finding appropriate MeSH subheadings is one way to narrow down the search to a specific aspect of the query topic. On the other hand, if we know what the specific aspect could be, we can search them in the document title or abstract. For example, if some resource gives all the possible roles genes can have on diseases in general, we can search those specific roles in the document title or abstract, instead of searching the word "role". UMLS semantic relations can be used in this situation.

The Unified Medical Language System (UMLS) has integrated more than 100 biomedical information resources, with more than 900,000 concepts (biomedical meanings). These concepts are categorized and related to each other by the UMLS Semantic Network. A semantic type in the Semantic Network is a category assigned to concepts based on their intrinsic and functional properties. For example, "kidney" is a "Body Part, Organ, or Organ Component" and "Sexuality" is a Behavior. The current release of the Semantic Network contains 135 semantic types. These semantic types denote physical objects, ideas, activities, biologic functions, anatomical structural, and chemicals.

The UMLS semantic relations can be grouped into five categories: physical (e.g., "Tissue" connected_to "Body Part,Organ, or Organ Component"), functional (e.g., "Virus" causes "Pathologic Function"), spatial (e.g., "Fungus" location_of "Enzyme"), temporal (e.g., "Genetic Function" co-occurs_with "Mental Process") or conceptual (e.g., "Disease or Syndrome" degree_of "Mental or Behavioral Dysfunction"). In our experiment, only functional and temporal semantic relations are used.

Step 1: Identify concepts and their semantic types in each topic template

For example, two concepts are mapped in topic template 2: "Role of gene" maps to concept "gene function". The semantic type of "gene function" is "genetic function"; "Disease" is a concept and its semantic type is “Disease or Syndrome”.

Step 2: Retrieve semantic relations between the semantic types from UMLS semantic network. Only temporal and functional relations are considered.
Continue the above example, the relations between “genetic function” and “Disease or Syndrome” in UMLS are:

Genetic Function	affects	Disease or Syndrome
Genetic Function	process_of	Disease or Syndrome
Genetic Function	result_of	Disease or Syndrome

“Affects”, “process_of”, and “result_of” are all functional relations.

Step 3: Extract the more specific relations in the definition of each semantic relation.
For example, “Affects” is defined in the UMLS as: “Produces a direct effect on. Implied here is the altering or influencing of an existing condition, state, situation, or entity. This includes has a role in, alters, influences, predisposes, catalyzes, stimulates, regulates, depresses, impedes, enhances, contributes to, leads to, and modifies.” By looking for the word “includes” in the definition, those more specific types of “Affects” are exacted: “has a role in”, “alters”, “ influences”, “predisposes”, “catalyzes”, “stimulates”, “ regulates”, “depresses”, “impedes”, “enhances”, “contributes to”, “leads to”, and “modifies”.

Step 4: Add the more specific relations into the original query.
They are ORed together and added as ordinary terms, not as concepts. For the above example, the added relations will be:

has a role in |alters|influences|predisposes|catalyzes|stimulates |regulates|depresses |impedes|enhances|

contributes to|leads to|modifies

2. Gene name disambiguation

The ambiguity of gene symbols is substantial, not only because one symbol may denote multiple genes but particularly because many symbols have other, non-gene meanings. For example, gene APC also has many non-gene meanings, such as “Antigen Presenting Cells”, or “Activated Protein C”, or “Argon Plasma Coagulation”. These two types of ambiguity provide a great challenge for information retrieval. Recent studies include [7] and [8]. [7] introduced a system, *Gpmarkup*, which automatically identifies gene/protein names in MEDLINE abstracts and links the names to their synonyms. Their system also disambiguates homonyms (two or more gene/protein names spelled alike but different in meaning) by mapping them to the full forms. [8] proposed an algorithm to disambiguate homonymous gene symbols by comparing their thesaurus descriptions and the textual contexts in which they occur (MEDLINE abstracts).

3. Psudeo feedback

Psudeo feedback has been proved to help query expansion. Those concepts extracted from the top ranked documents can be related and added to the concepts in the query. By relating and adding this kind of concepts to the query may help retrieve more relevant documents. Concepts extracted from the top ranked documents are related to the concepts in the query according to the MeSH hierarchy. The added concepts are either the hyponyms or hypernyms of the query concepts. Hyponyms are weighted the same as the original concept in the query. But the hypernyms are weighted less than the original concept, since they are more general.

Acknowledgments

This research is funded jointly by the National Library of Medicine and the National Institute of Mental Health. Thanks Professor Neil R. Smalheiser for helpful discussions.

References

[1] William Hersh. (2005) TREC 2005 Genomics Track Protocol. http://ir.ohsu.edu/genomics/2005protocol.html

[2] http://www.nlm.nih.gov/mesh/filelist.html

[3] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. (2004) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2005 Jan 1;33(Database issue):D54-8. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[4] http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#AutomaticTermMapping

[5] Kostoff RN, Block JA, Stump JA, Pfeil KM.(2004) Information content in Medline record fields. Int J Med Inform. 2004 Jun 30;73(6):515-27.

[6] S.E. Robertson, S. Walker (2000) Okapi/Keenbow at TREC-8. NIST Special Publication 500-246:The Eighth Text REtrieval Conference (TREC 8)

[7] Yu H, Hatzivassiloglou V, Rzhetsky A, Wilbur WJ.(2002) Automatically identifying gene/protein terms in MEDLINE abstracts. J Biomed Inform. 2002 Oct-Dec;35(5-6):322-30

[8] Schijvenaars BJ, Mons B, Weeber M, Schuemie MJ, van Mulligen EM, Wain HM, Kors JA.(2005) Thesaurus-based disambiguation of gene symbols. BMC Bioinformatics. 2005 Jun 16;6:149.