

# Relation-Based Document Retrieval for Biomedical Literature Databases\*

Xiaohua Zhou, Xiaohua Hu, Xia Lin, Hyoil Han, and Xiaodan Zhang

College of Information Science & Technology, Drexel University,  
3141 Chestnut Street, Philadelphia, PA 19104  
xiaohua.zhou@drexel.edu,  
{thu, xlin, hhan, xzhang}@cis.drexel.edu

**Abstract.** In this paper, we explore the direct use of relations in information retrieval for precision-focused biomedical literature search. A relation is defined as a pair of two concepts which are semantically and syntactically related to each other. Unlike the traditional term-based IR models, our model represents a document by a set of controlled concepts and their binary relations. Since document level co-occurrence of two concepts, in many cases, does not mean this document really addresses their relationships, the direct use of relation may improve the precision of very specific search, e.g. *searching documents that mention genes regulated by Smad4*. For this purpose, we develop a generic ontology-based approach to extract concepts and their relations; a prototyped IR system supporting relation-based search is then built for Medline abstract search. We then use this novel IR system to improve the retrieval result of all official runs in TREC-2004 Genomics Track. The experiment shows promising performance of relation-based IR. The mean of P@100 (the precision of top 100 documents) for all 50 topics is raised from 26.37 % (the P@100 of the best run is 42.10%) to 53.69% while the recall is kept at an acceptable level of 44.31%. The experiment also demonstrates the expressiveness of relations for the representation of genomic information needs.

## 1 Introduction

Precision and recall are two basic metrics measuring the performance of an Information Retrieval (IR) system. Often, high precision is at the cost of low recall, and vice versa. Nowadays, precision-focused searching is getting more and more attention most likely due to the following two reasons. First, in a lot of domain-specific search, such as searching the Medline, which collects 14 millions of biomedical abstracts published in more than 4600 journals, the professionals normally know what they need and their search queries are often very specific and only like to receive those documents which meet their specific query; thus, they do not expect a large number of documents. Second, the absolute number of returned relevant document is still large enough for most retrieval tasks even if the recall is low because of the exponentially increasing size of the document collection.

---

\* This research work is supported in part from the NSF Career grant (NSF IIS 0448023). NSF CCF 0514679 and the research grant from PA Dept of Health.

Traditional IR models often use a set of terms to index and search documents. A term might be a concept from a controlled vocabulary, or a word or a phrase in a natural language statement, or a thesaurus entry representing a set of synonymous terms [14]. Term-based indexing and searching is convenient for text processing. However, this mechanism might lose some useful information such as the correspondence between terms strongly addressed in the original documents. There are full of various explicitly asserted biological relationships in genomic and biomedical literature, e.g. protein interactions and disease complications; these biological relationships are exactly what scientists are interested in. Therefore, we hypothesize that the direct use of relationships would improve the precision of genomic information retrieval (GIR).

Term-based IR models have to use term co-occurrence to approximate relations because there are no direct relations available in their indices. However, the co-occurrence of two terms in a document, in many cases, does not mean this document really addresses their relationships, especially when the co-occurrence count is low (e.g. in abstract-based search such as PubMed). Thus, the precision would be compromised. We conducted a simple experiment that tried to retrieve documents addressing the interaction of *obesity* and *hypertension* from PubMed<sup>1</sup> by specifying the co-occurrence of term *hypertension* and *obesity* in abstract or title. We then took the top 100 abstracts for human relevance judgment. Unfortunately, as expected, only 33 of them were relevant.

*obesity [TIAB] AND hypertension [TIAB] AND hasabstract [text]  
AND ("1900"[PDAT] : "2005/03/08"[PDAT])*

**Fig. 1.** The query used to retrieve documents addressing the interaction of obesity and hypertension from PubMed. A ranked hit list of 6687 documents is returned.

In literature, there are volumes of work using term relationships to improve IR. However, their definition of the relationship and the motivation to use relationships are different from ours. Their relationships could be roughly classified into two classes. One is the co-occurrence relationship; the range for co-occurrence might be a document, a paragraph, a sentence, or a fix-sized sliding window [1, 2, 20]. The other is the general semantic relationship such as is-a, part-of and synonym [2]. Their applications of relationships in IR also fall into two categories. One line of work applies the correspondence between query terms and document terms into query expansion [1]. The other line of work uses the syntactic relationship between document terms to estimate a more accurate dependency document model such as bigram and trigram [5, 11]. The effect of the dependency model on IR is similar to that of using phrases instead of words as the indexing unit.

Our relation is defined as a pair of two concepts which are semantically and syntactically related to each other. The semantic constraint could be but not limited to general is-a, part-of and synonym. In most cases, they refer to domain-specific relationships. For GIR, the semantic relationships could be interaction, binding, affecting, producing, etc. The syntactic constraint is the explicit assertion of the binary relation between two concepts in a natural language statement. Many general (e.g. WordNet)

<sup>1</sup> <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>

or domain thesaurus (e.g. UMLS) already define lots of semantic relationships. But if the concepts of a relation are not syntactically related in the document they appear, we would not treat them as a relationship during both indexing phase and searching phase. Thus, our definition of relationship is stricter than that in the previous literature. Our motivation for relationship is also different from previous work. We directly use relationships in conjunction with concepts to index and search documents whereas previous work indirectly uses relationships for query expansion or dependency document model estimation.

The extraction of binary relations from text is a challenging task. We think this is one of the major reasons that no relation-based search engine is reported so far. The concept (term) extraction is the first step of the relation extraction. The methods for term extraction fall into two categories, with dictionary [13, 21] or without dictionary [9, 12, 17, 18]. The later is characterized by its high extracting speed and no reliance on dictionary and the capability of predicting new terms. However, it does not extract the meaning of a term. Thus, it does not fit for our application. Instead, we apply a dictionary-based approach [21] to the concept extraction. The majority of the literature use patterns learned by supervised approaches [12] or unsupervised approaches [7, 13], or coded by hand to identify binary relations in a natural language statement. We apply hand-coded patterns to the extraction of binary relations.

We finally develop a generic ontology-based approach to extract concepts and their binary relations. Based on that, we build a prototyped IR system supporting relation-based search for Medline abstracts. We use this novel IR system to improve the retrieval result of all official runs in TREC-04 Genomics Track. The experiment shows promising performance of relation-based IR. The mean of P@100 (the precision of top 100 documents) for all 50 topics is raised from 26.37 % (the P@100 of the best run is 42.10%) to 53.69% while the recall is kept at an acceptable level of 44.31%. The experiment also shows the expressiveness of relations for the representation of information needs, especially in the area of biomedical literature which are full of various biological relations.

The rest of the paper is organized as follows: Section 2 describes the representation of documents and queries. Section 3 presents a generic approach to the extraction of concepts and relations. Section 4 shows the experiment design and result. A short conclusion finishes the paper.

## 2 Representation of Document and Query

Traditional IR models a document and a query as a set of terms. A term might be a concept from a controlled vocabulary, or a word or a phrase in a natural language statement, or a thesaurus entry representing a set of synonymous terms. The different indexing units may produce slightly different performance for IR. But neither of them explicitly addresses the relation between terms, i.e. terms in a document are unstructured. Obviously, a document is full of various relations. For example, biomedical literatures contain a large number of biological relationships among gene, protein, mutation, disease, drug, etc. Intuitively, the incorporation of such knowledge (represented by relations) will help improve the precision of an IR system. For this purpose, we propose a relation-based document representation mechanism below.

## 2.1 Document Representation

In the relation-based IR model, we represent a document by a set of concepts from UMLS and their binary relations as shown in Figure 2. We use controlled concepts rather than words in natural language to index the documents because of the characteristics of the GIR. In genomic-related literature, a term is often comprised of multiple words; the word-based unigram IR model might lose the semantics of the term. Meanwhile, severe synonym and polysemy problem in GIR might cause trouble while an IR system tries to match query terms with indexing terms according to their names instead of meanings [15, 19]. A UMLS concept is a meaning with a unique ID representing a set of synonymous terms. Thus, the introduction of UMLS concept for indexing may relieve the two above-mentioned problems.

|   |                |
|---|----------------|
| <b>Terms (CUI, Name, Semantic Type, Frequency)</b>          |                |
| T1 (C0003818, arsenic, Hazardous or Poisonous Substance, 9) |                |
| T2 (C0870082, hyperkeratosis, Disease or Syndrome, 4)       |                |
| T3 (C1333356, XPD, Gene, 6)                                 |                |
| T4 (C0007114, skin cancer, Neoplastic Process, 1)           |                |
| T5 (C0012899, DNA repair, Genetic Function, 3)              |                |
| T6 (C0241105, hyperkeratotic skin lesion, Finding, 2)       |                |
| T7 (C0936225, inorganic arsenic, Inorganic Chemical, 1)     |                |
| .....   |                |
| <b>Relations (First Concept, Second Concept, Frequency)</b> |                |
| R1 (T1, T3, 3)  | R2 (T2, T4, 1) |
| R3 (T2, T5, 2)  | R4 (T2, T3, 2) |
| R5 (T4, T5, 1)  | R6 (T3, T4, 1) |
| .....   |                |

**Fig. 2.** A real example of document representation. The document (PMID: 12749816) can be found through PubMed. CUI is the unique ID of a concept in UMLS<sup>2</sup>.

However, we keep term names in the index because term names do provide additional information for IR. For example, in the experiment of TREC 2004 Genomics Track (see Section 2.2 and Section 4), we use term names to decide if a term (protein) belongs to certain protein family. Also, we record the semantic type of a term, the category a term belongs to. The semantic type is also useful to express information needs (see Section 2.2).

A relation is defined as a pair of two concepts which are semantically and syntactically related to each other. We extract all such concept pairs in a document and record their frequency. For the simplicity, the relation in our model is undirected.

## 2.2 Query Representation

The query representation is often subject to the mechanism of document representation. Under traditional term-based IR model, we often use term vector or term-based Boolean expression to represent information needs. In this section, we will first

<sup>2</sup> <http://www.nlm.nih.gov/research/umls/>

briefly introduce the syntax of relation-based Boolean expression and then demonstrate the effectiveness of this query representation mechanism by the examples from TREC 2004 Genomics Track.

Two types of predicates denoted by concept (T) and relation (R) are available to build Boolean expression. A concept can be specified by any combination of its name (STR), unique ID (CUI), and semantic type (TUI). All predicates can be combined by AND or OR operator. Here, we use the ad hoc retrieval topics in TREC 2004 Genomics Track<sup>3</sup> to illustrate how to use relation-based Boolean expression to represent user information needs.

**Topic #1:** Ferroportin-1 in humans

**Query:** T (CUI=C0915115)

**Notes:** C0915115 is the concept ID of *Ferroportin-1* in the dictionary of UMLS (Unified Medical Language System). All concepts IDs in this paper are based on UMLS.

**Topic #12:** Genes regulated by Smad4

**Query:** R (CUI<sub>1</sub>=C0694891 and TUI<sub>2</sub>=T028)

**Notes:** C0694891 is the concept ID of *Smad4* and T028 stands for the semantic type if *Gene*. Because a relation is undirected, the query should contain the symmetric predicate R (CUI<sub>2</sub>=C0694891 and TUI<sub>1</sub>=T028). However, for the simplicity, we let the IR system automatically generate the symmetric predicate R.

**Topic #14:** Expression or Regulation of TGF $\beta$  in HNSCC cancers

**Query:** R (CUI<sub>1</sub>=C1515406 and CUI<sub>2</sub>=C1168401)

**Notes:** C1515406 is the concept ID of *TGF $\beta$*  and C1168401 is the concept ID of *HNSCC*

**Topic #30:** Regulatory targets of the Nkx gene family members

**Query:** R (STR<sub>1</sub> like nkx% and TUI<sub>1</sub>=T028 and TUI<sub>2</sub>=T028)

**Notes:** we assume a term with its name beginning with nkx and with semantic type of gene is the member of *Nkx gene family*.

We can see that relation-based Boolean expression is neat and powerful to express user information needs from above examples. In topic #1, we simply use one T predicate though Ferroportin-1 has lots of synonyms. In topic #12 and #30, we use one R predicate in conjunction with semantic types to express a question-answering type information need that is very difficult to be represented by term vector or term-based Boolean expression.

### 3 Extraction of Concepts and Relations

In this section, we propose a generic ontology-based approach to the extraction of concepts and relations. As shown in Figure 3, we first extract term names using domain ontology in conjunction with part of speech patterns [21]; then use surrounding words to narrow down the meaning of the extracted term, i.e. identifying the concept the term refers to in the context. Finally we employ several heuristic approaches to the extraction of binary relations.

<sup>3</sup> <http://trec.nist.gov/data/genomics/04.adhoc.topics.txt>

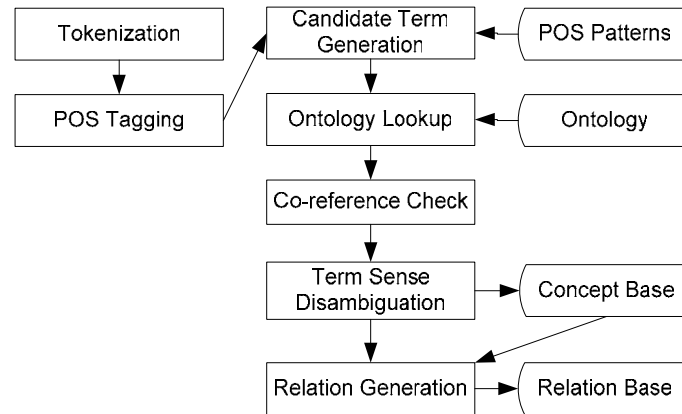


Fig. 3. The architecture of the concept and relation extraction system

### 3.1 Extraction of Terms

There are volumes of work on the topic of term extraction from biomedical literatures. Most of them use either hand-created rules or machine-learned rules to extract terms from text. However, neither of them extracts the meaning of a term that is important to our document representation model. For instance, the information extraction (IE) system may tell you that *Ferroportin-1* is a protein but not tell you what protein it is. For this reason, we implement a generic ontology-based approach [21] that identifies not only the semantic type of the term, but also its possible meanings. This approach begins with part of speech (POS) tagging, then generates candidate terms using POS patterns, and finally determines if it is a term by looking up the ontology.

In this particular project, we take UMLS as the domain ontology. UMLS is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms in the area of biomedicine and health. The Metathesaurus of UMLS is organized by concept or meaning of terms and provides their various names (synonyms), and the relationships among them. By checking with the synonym table, we can easily determine if the candidate (generated by POS patterns listed in Table 1) is a term and retrieve possible meanings if yes.

**Table 1.** Part of Speech Patterns and Examples. NN, NUM, and JJ denote noun, number, and adjective, respectively. All article, preposition, and conjunction words will be removed from the original text before pattern matching.

| Part of Speech Pattern | Examples                |
|------------------------|-------------------------|
| NN NN NN               | Cancer of Head and Neck |
| NN NUM NN              | DO 1 Antibody           |
| JJ NN NN               | High Blood Pressure     |
| NN NN                  | DNA Repair              |
| NN NUM                 | Ferroportin 1           |
| JJ NN                  | Sleeping Beauty         |
| NN                     | FancD2                  |

A term sometimes appears in the form of a pronoun such as *it* or its abbreviation. It is then necessary to figure out what the pronoun or the abbreviation refers to in the local context. We then develop a simple heuristic approach to handle abbreviations and implement a light method [3] to solve pronominal references.

### 3.2 Term Sense Disambiguation

Using the approach proposed by [21], we may extract more than one meaning for a term. For example, *Ferroportin-1* has two meanings in UMLS (*C0915115: metal transporting protein 1; C1452618: Slc40a1 protein, mouse*). Thus, we need a sense disambiguation component to further clarify the meaning the term refers to in the context.

Inspired by the finding that the ambiguity of many UMLS terms in text is caused by the use of short name, abbreviation, or partial name, we develop an unsupervised term sense disambiguation approach adapted from Lesk's word sense disambiguation (WSD) approach. The Lesk's WSD approach basically tags sense by maximizing the number of common words between the definition of candidate senses and the surrounding words of the target [10]. Different from Lesk's approach, our approach first use surrounding words (3 words in the left side of the target and 3 word in the right side of the target) to narrow down sense candidates. If there is still more than one sense left, we then score each candidate. In Lesk's approach, any word in any sense has same weight. Obviously it is not a good assumption for term sense disambiguation. Instead, we borrow the idea from term weighting research and use TF\*IDF to score the importance of a word for a sense [8]. Then the final formula for sense tagging is:

$$S = \arg \max_j \sum_i IDF_i \times TF_{ij} = \arg \max_j \sum_i \log \frac{N}{n_i} \times \frac{F_{ij}}{F_j}$$

Where:

$N$  is the number of senses in dictionary

$n_i$  is the number of senses containing  $Word_i$

$F_{ij}$  is the occurrence of  $Word_i$  in various names of  $Sense_j$

$F_j$  is the total occurrence of words in various names of  $Sense_j$

### 3.3 Extraction of Relations

A relation is defined as a pair of two concepts which are semantically and syntactically related to each other. If there is a pre-defined relation between the semantic types of two concepts in the domain ontology, these two concepts are simply viewed as semantically related. However, the judgment of syntactic relation between two concepts is difficult. We propose a heuristic approach for syntactic relation judgment.

The extraction of biological relationships is a hot topic in the area of information extraction. The essence of this line of work is to generalize the syntactic rules for certain types of relations in a supervised or unsupervised manner. However, there are two major problems when applying these methods to extract biological relations for our IR indexing. First, the indexing component of our IR system is interested in various biological relationships. But most of these reported extracting methods are merely

tested on protein-protein interactions. Second, the recall of these extracting methods seems to be low for IR use. For example, the IE system reported by [13] only extracts 53 relationships with 43 correct from 1,000 Medline abstracts containing the keyword “protein interaction”. Instead, we develop a simple but effective heuristic approach that first uses clause level co-occurrence to generate concept-pair candidates and then apply a set of rules to filter out some candidates. This approach is able to identify various biological relationships with high recall and good precision for IR use.

Term co-occurrence is frequently used to determine if two terms are connected in graph-based data mining. Some work takes any pair of two words in a sentence as a relation [16]. However, as reported by [4], sentences in Medline abstracts are often very long and complex. Thus, if we follow the strategy of [16], many noisy relations may be introduced. Instead, we use a clause as the boundary of a relation because concepts within a clause are more cohesive than within a sentence in general. We implement a light approach that basically uses comma and a set of conjunction words (including *although*, *because*, *but*, *if*, *that*, *though*, *when*, *whether*, *while* and so on) to split a complex sentence into one main clause and several subordinating clauses. In example 1, there are three terms underlined and one relation (*obesity* and *periodontal disease*). The term *epidemiological study* has no relation with any of the other two terms because it is in a separate clause.

**Rule for relation:** *If two concepts are co-occurred within a clause, but are not coordinating components, and their semantic types are related to each other in domain ontology, this concept pair is identified as a binary relation.*

**Example 1:** *A recent epidemiological study revealed that obesity is an independent risk factor for periodontal disease.*

**Example 2:** *Diabetes is associated with many metabolic disorders including insulin resistance, dyslipidemia, hypertension and atherosclerosis.*

Also, Ding et al. [4] pointed out that coordinating was a frequently occurred phenomenon in biomedical documents and interactions (relations) between coordinating components was rare in Medline abstract. Thus, in example 2, *diabetes* has relations with remaining four concepts respectively. But *insulin resistance*, *dyslipidemia*, *hypertension*, and *atherosclerosis* don't have relations with each other because they are coordinating components.

In short, we consider a concept pair a binary relation if these two concepts are co-occurred within a clause, but are not coordinating components, and their semantic types are related to each other in the domain ontology.

## 4 Experiment

In this section, we discuss the experiment design and the search engine and document collection used for experiment. Then we analyze the experiment result and compare the performance of proposed relation-based IR model with other work.

#### 4.1 Search Engine and Collection for Experiment

To our best knowledge, no search engines support relation-based search so far. For this reason, we developed a prototyped IR system supporting relation-based Boolean search. We implemented conceptual document representation in Figure 2 with a DB2 database. When a query represented by relation-based Boolean expression (see Section 2.2) is submitted, the system automatically converts the Boolean expression to ANSI SQL statement and submits the SQL statement to the DB2 system. The prototyped IR system is function-limited. It does not support document ranking, but simply returns documents that satisfy all predicates specified in the query.

We use the collection of TREC 2004 Genomics Track in our experiment. The document collection is a 10-year subset (1994-2003, 4.6 million documents) of the MEDLINE bibliographic database of the biomedical literature that can be searched by PubMed. Relevance judgments were done using the conventional "pooling method" whereby a fixed number of top-ranking documents from each official run were pooled and provided to an individual for relevance judgment. The pools were built from the top-precedence run from each of the 27 groups. They took the top 75 documents for each topic and eliminated the duplicates to create a single pool for each topic. The average pool size (average number of documents judged per topic) was 976, with a range of 476-1450. Based on the human relevance judgment, the performance of each official run could be evaluated (All facts and evaluation result of TREC-04 Genomics Track in Section 5 are from [6]).

Since our goal is to see whether our relation-based IR methods can further improve TREC 2004 participants' retrieval results, we build our search engine on top of search engines participated in TREC 2004. For this, we take the documents in pools for each topic and eliminate repeated documents across topics to create a single pool for our experiment use. The indexing and searching of our prototyped IR system is based on this mini-pool containing 42, 255 documents.

#### 4.2 Experiment Design

Our goal is to build a precision-focused IR system. The major research question of this paper is *if relation-based IR outperforms term-based IR in terms of precision*. Because the current prototyped system does not support ranking, we compare overall precision (the precision of all retrieved documents) of our run with P@100 of all runs participated in TREC 2004 Genomics Track. Our run retrieved 125 documents on average. Thus, the comparison is fair to runs in TREC-04.

The hypothesis that relation-based IR outperforms term-based IR in terms of precision is actually based on the assumption that explicit assertion of term relation is more useful than document level term co-occurrence when judging if a document addresses certain relationship. To test the truth of this assumption, we study if the query  $R(t_1, t_2)$  provides higher precision than the query  $T(t_1)$  and  $T(t_2)$  in our experiment.

We are also interested in the recall of relation-based IR though it is not our focus. On one hand, the use of relation will lower the recall because the number of documents returned by  $R(t_1, t_2)$  is always equal or less than by  $T(t_1)$  and  $T(t_2)$ . On the other hand, the use of concept instead of term name well solves the synonym problem; thus it may increase the recall. So we will study the effect of use of concept and relation on the recall of IR.

### 4.3 Analysis of Experiment Result

Our run retrieves 124.80 documents on average and achieves 53.69% overall precision and 44.31% overall recall (see Table 5). Because our prototyped system does not support ranking, we compare our overall precision with P@100 of TREC 2004 Genomics Track. This comparison is fair to runs of TREC since our system retrieves more than 100 documents on average.

We first compare the precision on 50 individual topics. Except for topic 16, the precision of ours outperforms P@100 of TREC on all other 49 topics as shown in Fig. 4. Then we compare the precision of our run with P@100 of all official runs in TREC. As shown in Table 2, the precision of our run (53.69%) is significantly higher than P@100 of the top 3 runs and the mean of all official runs (26.37%). It is worth noting that we can not say that the precision of our IR system is better than that of other IR systems because our search is based on the returns of all other IR systems. But the experiment result really tells us that the relation based model is very promising for precision-focused IR because it significantly improves the precision of other IR systems.

For seven topics that use a single R predicate like  $R(CUI_1=A \text{ and } CUI_2=B)$ , we further change the Boolean expression to  $T(CUI=A)$  and  $T(CUI=B)$  and search again. As expected, the precision is lowered while the recall is improved (see Table 3). That is, the binary relation provides higher precision than document level term co-occurrence when retrieving documents addressing certain relationships. This is the foundation of the claim of the whole paper that relation-based IR model contributes higher precision to domain-specific research than term-based IR models.

The concept-based search can raise the recall of IR especially when a term has lots of synonyms because all synonyms share one concept ID. To test this hypothesis, we change seven single T predicate searches listed in Table 4 to term-based searches. As expected, the recall of topic 1 and 35 is significantly lowered because both of them have many synonyms.

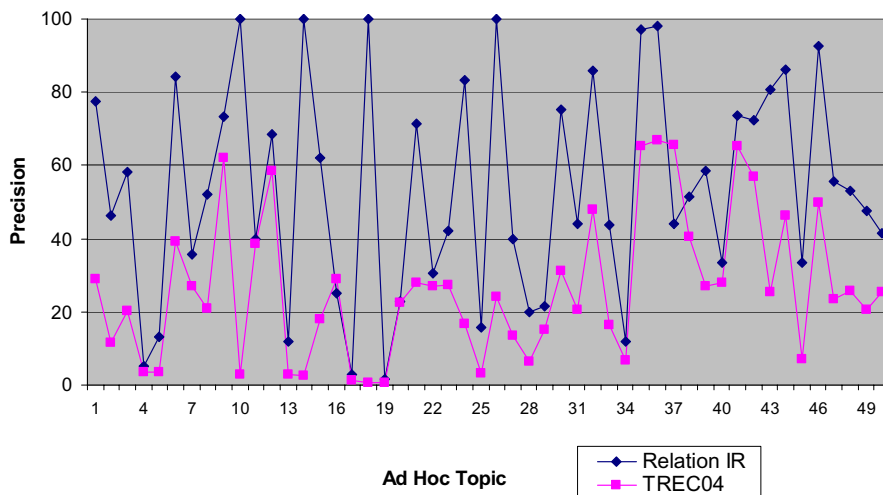


Fig. 4. The comparison of the overall precision of our relation-based IR system with the mean P@100 of all official runs in TREC 2004 Genomic Track on 50 ad hoc retrieval topics

**Table 2.** The comparison of the precision of our run with official runs participated in TREC 2004 Genomics Track. Runs in TREC are sorted by Mean Average Precision (MAP) [6]. Because our retrieval is not ranked, MAP and P@10 are not available; the P@100 of our run is actually the overall precision.

| Run                     | MAP   | P@10  | P@100 |
|-------------------------|-------|-------|-------|
| Relation IR (Our Run)   | N/A   | N/A   | 53.69 |
| pllsgen4a2 (the best)   | 40.75 | 60.04 | 41.96 |
| uwntDg04tn (the second) | 38.67 | 62.40 | 42.10 |
| pllsgen4a1 (the third)  | 36.89 | 57.00 | 39.36 |
| edinauto5 (the worst)   | 0.12  | 0.36  | 1.3   |
| Mean@TREC04             | 21.72 | 42.69 | 26.37 |

**Table 3.** The comparison of the use of relation and concept co-occurrence in IR

| Topic | $R(t_1, t_2)$ |       | $T(t_1) \text{ and } T(t_2)$ |       | P@100<br>TREC04 (%) |
|-------|---------------|-------|------------------------------|-------|---------------------|
|       | P (%)         | R (%) | P (%)                        | R (%) |                     |
| 7     | 35.71         | 8.70  | 24.62                        | 27.83 | 27.04               |
| 8     | 52.00         | 8.07  | 41.05                        | 24.22 | 20.94               |
| 13    | 12.00         | 12.50 | 8.77                         | 20.83 | 2.74                |
| 14    | 100.00        | 23.81 | 80.00                        | 23.81 | 2.70                |
| 15    | 61.90         | 14.44 | 48.08                        | 27.78 | 18.00               |
| 21    | 71.43         | 18.75 | 52.83                        | 35.00 | 27.96               |
| 22    | 30.52         | 44.76 | 25.14                        | 65.71 | 27.09               |

**Table 4.** The comparison of concept-based search and term-based search

| Topic | Name B          | T(CUI=A) |       | T (STR like %B%) |       | T(STR=B) |       |
|-------|-----------------|----------|-------|------------------|-------|----------|-------|
|       |                 | P (%)    | R (%) | P (%)            | R (%) | P (%)    | R (%) |
| 1     | Ferroportin     | 77.59    | 56.96 | 84.62            | 41.77 | 88.46    | 29.11 |
| 6     | FancD2          | 84.09    | 39.36 | 84.09            | 39.36 | 85.29    | 30.85 |
| 9     | mutY            | 73.38    | 98.26 | 81.75            | 97.39 | 81.48    | 95.65 |
| 35    | WD40            | 97.16    | 63.10 | 99.28            | 50.55 | 98.28    | 21.03 |
| 36    | RAB3A           | 98.10    | 81.50 | 98.10            | 81.50 | 98.53    | 79.13 |
| 43    | Sleeping Beauty | 80.56    | 14.87 | 77.42            | 12.31 | 77.42    | 12.31 |
| 46    | RSK2            | 92.59    | 12.69 | 82.76            | 12.18 | 89.47    | 8.63  |

## 5 Conclusions and Future Work

In this paper, we proposed a novel relation-based information retrieval approach for biomedical literature search. Unlike traditional term-based IR models that use terms to index and search documents, our relation model uses controlled concepts and their binary relations to index and search documents. Because explicitly asserted biological relationships are exactly what scientists are interested in, the direct use of relation for document indexing and searching may improve the precision of genomic information retrieval. The experiment on the collection of TREC 2004 Genomics Track successfully tested this hypothesis. Besides, we could draw another three conclusions from the experiment:

- An explicitly asserted relation in text is a stronger indicator of a document that addresses a binary relation than the document level concepts co-occurrence.
- Concept-based search will bring higher recall than term-based search especially when a searching term has many synonyms.
- Relation-based Boolean expression is powerful and effective to express genomic information needs.

For future work, we will develop a ranking algorithm for relation-based IR and implement a full-functioned search engine supporting relation-based searching. We will also take effort on the extraction of concepts and relations that would further improve the performance of the relation-based search.

## References

1. Bai, J., Song, D., Bruza, P., Nie, J.Y., and Cao, G., "Query Expansion Using Term Relationships in Language Models for Information Retrieval", *In Proceedings of the ACM 14th Conference on Information and Knowledge Management (CIKM)*, November 2005, Bremen, Germany.
2. Cao, G., Nie, J.Y., and Bai, J., "Integrating Word Relationships into Language Models", *Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 2005, pp. 298 - 305
3. Dimitrov, M., Bontcheva, K., Cunningham, H., and Maynard, D., "A Light-weight Approach to Coreference Resolution for Named Entities in Text", *Proceedings of the Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, Lisbon, 2002.
4. Ding, J., Berleant, D., Xu, J., and Fulmer, A.W., "Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser", *In the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03)*, 2003.
5. Gao, J., Nie, J.Y., Wu, G. and Cao G., "Dependency Language Model for Information Retrieval", *Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 2004, pp. 170 - 177
6. Hersh W, et al. "TREC 2004 Genomics Track Overview", The thirteenth Text Retrieval Conference, 2004.
7. Hu, X., Yoo, I., Song, I.Y., Song, M., Han, J., and Lechner, M., "Extracting and Mining Protein-Protein Interaction Network from Biomedical Literature", *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2004.
8. Jones, K.S., "Exhaustivity and specificity", *Journal of Documentation*, 1972, Vol. 28, pp.11-21.
9. Kim, J.T. and Moldovan, D.I., "Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction", *IEEE Transactions on Knowledge and Data Engineering*, 1995, 7(5), pp. 713-724.
10. Lesk, M., "Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream Cone", *Proceedings of the SIGDOC'86 Conference*, ACM, 1986.
11. Miller, D., Leek, T., and Schwartz M.R., "A Hidden Markov Model Information Retrieval System", *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, 1999, pp 214-221.

12. Mooney, R. J. and Bunescu, R. "Mining Knowledge from Text Using Information Extraction", *SIGKDD Explorations* (special issue on Text Mining and Natural Language Processing), 7, 1 (2005), pp. 3-10.
13. Palakal, M., Stephens, M.; Mukhopadhyay, S., Raje, R., Rhodes, S., "A multi-level text mining method to extract biological relationships" , *Proceedings of the IEEE Computer Society Bioinformatics Conference (CBS2002)*, 14-16 Aug. 2002 Page(s):97 – 108
14. Salton, G. and Buckley, C., "Improving retrieval performance by relevance feedback", *Journal of the American Society for Information Science*, 1990, vol. 41, pp. 288-97
15. Sanderson, M. 1994, "Word sense disambiguation and information retrieval", *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, p.142-151, July 03-06, 1994, Dublin, Ireland.
16. Schenker, A., Last, M., Bunke, H., and Kandel, A., "Clustering of Web Documents Using a Graph Model", *In A. Antonacopoulos & J. Hu (Eds.), Web Document Analysis: Challenges and Opportunities*, 2003.
17. Soderland, S., Fisher, D., Aseltine, J., and Lehnert, W., "CRYSTAL: Inducing a Conceptual Dictionary", *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995, pp. 1314-1319.
18. Soderland, S., "Learning Information Extraction rules for Semi-structured and free text", *Machine Learning*, Vol. 34, 1998, pp. 233-272.
19. Stokoe, C. and Tait, J. I. 2004. Towards a Sense Based Document Representation for Information Retrieval, in *Proceedings of the Twelfth Text REtrieval Conference (TREC)*, Gaithersburg M.D.
20. van Rijsbergen, C.J., "A theoretical basis for the use of cooccurrence data in information retrieval", *Journal of Documentation*, 1977, 33(2), pp 106-119.
21. Zhou, X., Han, H., Chankai, I., Prestrud, A., and Brooks, A., "Converting Semi-structured Clinical Medical Records into Information and Knowledge", *Proceeding of The International Workshop on Biomedical Data Engineering (BMDE) in conjunction with the 21st International Conference on Data Engineering (ICDE)*, Tokyo, Japan, April 5-8, 2005.