Combining Semantics, Context, and Statistical Evidence in Genomics Literature Search

Jay Urbain
Information Retrieval Laboratory
Illinois Institute of Technology
Chicago, IL
urbajay@iit.edu

Nazli Goharian
Information Retrieval Laboratory
Illinois Institute of Technology
Chicago, IL
goharian@iit.edu

Ophir Frieder[1]
Computer Science Department
Georgetown University
Washington, DC
ophir@cs.georgetown.edu

***Abstract*—We present an information retrieval model for combining evidence from concept-based semantics, term statistics, and context for improving search precision of genomics literature by accurately identifying concise, variable length passages of text to answer a user query.**

The system combines a dimensional data model for indexing scientific literature at multiple levels of document structure and context with a rule-based query processing algorithm. The query processing algorithm uses an iterative information extraction technique to identify query concepts, and a retrieval function for systematically combining concepts with term statistics at multiple levels of context. We define context by variable length passages of text and different levels of document lexical structure including terms, sentences, paragraphs, and entire documents.

Our results demonstrate improved search results in the presence of varying levels of semantic evidence, and higher performance using retrieval functions that combine document as well as sentence and passage level information versus using document, sentence or passage level information alone.

Initial results are promising. When ranking documents based on the most relevant extracted passages, the results exceed the state-of-the-art by 13.89% as assessed by the TREC 2005 Genomics track collection of 4.5 million MEDLINE citations.

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval-search process; H.3.1 [Information Storage and Retrieval]: Context Analysis and Indexing-linguistic processing; I.2.7 [Artificial Intelligence]: Natural Language Processing-text analysis.

I. INTRODUCTION

Accurate retrieval of information from genomics literature is a key component in experiments to identify new genes, diseases, and other biological processes that require further investigation [1].

Information retrieval in this domain is challenging due to the wide variation of synonymous terms, acronyms, and morphological variants used for identifying the same biological concepts. In addition, acronyms frequently have multiple meanings (polysemy) and require contextual clues for accurate disambiguation. For example, the terms "bovine spongiform encephalopathy", "BSE", and "Mad Cow Disease" are all different terms representing the same named entity or concept. Search terms also have much higher relevance when matched against document terms when occurring within the local context of a phrase, sentence, or passage of text. An acronym like "IP" could represent "immunoprecipitant" or "ischemic precondition." In this case, context captured at the paragraph or document level where an acronym is defined can help disambiguate its meaning.

Databases from the National Center for Biotechnology Information (NCBI) and other sources can be helpful in providing semantic evidence supporting identification and extraction of named biological entities [2]. However, it is important to recognize that no knowledge source can fully capture the complexities of human language let alone be fully up-to-date with the dynamic vocabulary of an evolving science. In most cases, there are varying levels of semantic evidence which can make accurate identification of biological concepts difficult. In these cases, optimal retrieval solutions need to integrate additional sources of evidence including identification of key phrases and terms within context and leverage traditional probabilistic measures of relevance.

We propose that effective search requires a systematic approach for combining semantic, contextual, and statistical evidence. Our approach relies on an indexing model that supports search of single and multi-word terms to support identification of concept term variants, search at different levels of document structure for identifying terms within context, and integration of external knowledge sources to aid in the identification and extraction of named biological entities and related synonymous terms.

We first describe our indexing model, followed by the indexing process, query processing, our methods, results, and a discussion of related work. For an introduction to information retrieval concepts refer to Grossman and Frieder [3].

[1]*Ophir Frieder is on leave from the Illinois Institute of Technology.*

II. INDEXING MODEL

Paragraphs, sentences, and terms, representing complete topics, thoughts, and units of meaning respectively, provide a logical breakdown of document lexical structure into finer levels of meaning and context.

We seek to capture these hierarchical relationships of document structure within a search index based on a dimensional data model. As shown in Figure 1, the dimensional index has a *dimension* table for each level of document structure (document, paragraph, sentence, term) and one *central fact* table or *postinglist*. The *postinglist* represents a single mapping table, containing foreign key fields that map the relations between all dimensions. The "grain", i.e., the smallest non-divisible element of the database, is the individual word. Sentences aggregate words in sequence by position, paragraphs aggregate sentences, and documents aggregate paragraphs. In the data warehousing literature, this model is refered to as a *star schema* [4,5].

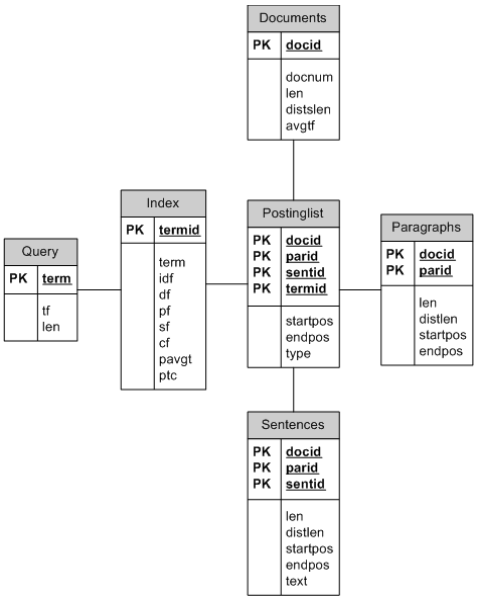

Figure 1. Search index based on dimensional model.

Term attributes include a term's position within a sentence, textual representation, as well as term and morphological variants.

The dimensional indexing model can be extended to include additional dimensions, and allows for efficient formulation of SQL search queries. By indexing each individual word, queries can be developed for searching single- and multi-word terms, and term statistics can be aggregated over different levels of document structure.

III. INDEXING PROCESS

The indexing process illustrated in Figure 2 includes:

1) *Lexical Partitioning*: Documents are parsed into sections (title, abstract, body text), and paragraphs. Paragraphs are parsed into sentences.

2) *Tokenization*: Acronyms and their long-forms are identified during indexing using the Schwartz and Hearst algorithm [6]. A long-short form would include "immuno deficiency enzyme (IDE)", and a short-long form would include "IDE (immuno deficiency enzyme)". The algorithm works backwards through the long form text and attempts to identify corresponding letters in the acronym. Acronyms and their long-forms are added to an acronym table to help with disambiguation. Long-form variants are added to the same indexing location as acronyms during indexing (and vice versa). This technique proved highly effective for disambiguating acronyms, and being able to identify and extract passages when searching for either the short- or long-form of an entity.

Sentence terms are tokenized, stop words removed, and lexical variants are generated [7]. Porter stemming [8] is used on each token with the following exceptions: gene names (as defined by the Entrez Gene database); all upper case, mixed case, alpha-numeric terms; and non-gene terms that would become a gene name after being stemmed. Small "s" is also stripped from all upper-case terms.

3) *Indexing*: Each term along with its long-form expansion and lexical variants are stored in the index with the same positional information.

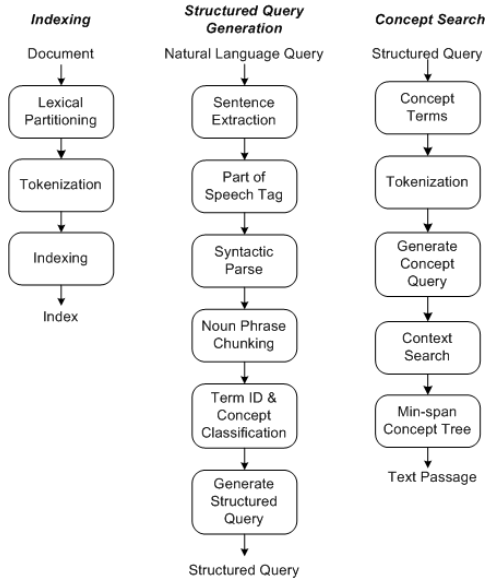

Figure 2. Process models

IV. QUERY PROCESSING

Structured query generation shown in Figure 2 is illustrated with the following query: "*Provide information about the role of the gene PRNP (prion protein) in the disease Mad Cow Disease"*.

1. Sentences are extracted, and acronyms and their long-forms are identified: PRNP (PRioN Protein).

2. Part-of-speed tagging is performed using our 2nd order statistical Hidden Markov Model tagger: … *role_NN of_II the_DD gene_NN PRNP_NN (_(prion_NN protein_NN)_) in_II the_DD disease_NN Mad_NN Cow_NN Disease_NN.*

3. Stop and function words are removed from further processing.

4. Candidate entities are identified by locating non-recursive noun phrases ("noun chunks"): *[gene PRNP], [prion protein], [Mad_NN Cow_NN Disease_NN]*.

5. Candidate entities are verified in the index, and resolved using the UMLS Metathesaurus®, OMIM™ (Online Mendelian Interface to Man), MeSH (Medical Subject Headings), and Entrez Gene databases. If an entity is successfully resolved, all synonyms and one level of hyponyms, i.e., child terms, are identified.

Prior to including synonyms as a concept term variant, its level of ambiguity is determined. If the synonym is considered ambiguous it is not included. We consider a term ambiguous if either of the following tests is met:

1. The synonym's normalized inverse document frequency (NIDF) is < 0.1. Where NIDF is the IDF=*log (N/df) normalized to between 0 and 1*.

2. The synonym correlates with the correct long-form in less than 50% of all instances within the acronym table

Resolved concepts and corresponding synonyms are shown in Table 1.

TABLE 1. BIOLOGICAL ENTITY RESOLUTION

Resolved concepts	Synonyms
[Encephalopathy, Bovine Spongiform]	[Mad Cow Disease] [MCD] [BSE] [Creutzfeldt-Jakob disease] [CJD]
[PRNP gene]	[prion protein] [prnp]

Search can be performed within the context of an individual term/phrase, sentence, paragraph, or document. In this study, we combine the search results of document, passage, and sentence retrieval. We first perform document and paragraph-level searches using the probabilistic BM25 (1) retrieval function [9] implemented in standard SQL [10].

BM25: (1)

$$\sum_{wq} \ln\left(\frac{N-df+0.5}{df+0.5}\right)\left(\frac{(k_1+1)*tf_d}{k1*(1-b)+b*(\frac{docLen}{avgDocLen})+tfd}\right)\left(\frac{(k_3+1)*tfq}{k_3+tf_q}\right)$$

Note: We used k1=1.4, k2=0, k3=7, and b=0.75.

Next, using the top 2000 paragraphs we perform a concept search as follows:

1. The position of all term variants of each concept is retrieved from the dimensional index by paragraph.

2. A concept graph is constructed by creating an adjacency list using each concept term as a vertex.

3. A minimum-spanning tree is constructed from the adjacency list by determining the maximum number of distinct concepts within the shortest lexical distance. Distance measurements are weighted such that terms within a lexical unit, e.g., a sentence, are always closer than terms in separate units.

4. Finally, the passage boundary based on the first and last occurrences of distinct concepts is expanded out to include sentence boundaries.

Passage level concept search is further illustrated with the following query: *"Exact reactions that take place when you do glutathione S-transferase (GST) cleavage during affinity chromatography"*.

First, the following concepts and term variants (shown in stemmed form) are identified:

- *Cleavage:* [[cleavag], [merogenesi], [cytokinesi]]
- *Affinity purification*: [affin, purif], [affin, chromatographi]]
- *Glutathione S-transferase*: [[glutathion, s, transferase], [gst]]

Second, the index is searched for all term variants of each concept. The following query searches for the concept term variant "*affinity, chromatography*":

```
select i1.term as term1, i2.term as term2, p1.docid, p1.parid,
       p1.sentid, p1.startpos, p1.endpos
from invertedindex i1, invertedindex i2, postinglist p1, postinglist p2
where i1.term='affin' and i2.term='chromatographi'
and i1.termid=p1.termid and i2.termid=p2.termid
and p1.docid=p2.docid and p1.parid=p2.parid
and p1.sentid=p2.sentid and abs(p2.seq-p1.seq)<=2;
```

Third, passages are identified: *"**affinity chromatography**, and purified Mce1A and Mce1E, free of the fusion partner, were recovered following specific proteolytic **cleavage** of the **GST**"*

Finally, passages are expanded to sentence boundaries: *"The fusion proteins were purified to near homogeneity by **affinity chromatography**, and purified Mce1A and Mce1E, free of the fusion partner, were recovered following specific proteolytic **cleavage** of the **GST** portion by thrombin protease."*

Passages are first ranked by the distinct number of concepts within the passage. For passages with the same number of concepts, the passages are further ranked by a query term density match (QTM) measurement we devised and used successfully for the 2006 TREC Genomics track [10]. QTM (2) assigns a "term match" score to each sentence within a passage by summing the NIDF of each *distinct* match of a query term at the sentence level.

$$QTM = \sum_{i=1}^{n} NIDF(i) \quad (2)$$

Sentences are ranked using the same technique as passages, except the passage length of one sentence.

Finally, a linear weighted sum (3) is used to combine the normalized BM25 document similarity coefficient with the normalized passage and sentence similarity coefficients (SC):

$$SC_{composite} = w_1SC_1 + w_2SC_2 + \ldots + w_nSC_n \quad (3)$$

V. METHODS

We evaluated our system on the TREC-2005 Genomics ad-hoc retrieval task which uses a corpus of 4,591,008 MEDLINE citations (~15GB) and 49 query topics drawn from the information needs of molecular biology researchers [11]. Each Medline citation includes an article title, medical subject headings (MESH), and typically includes an abstract.

To establish a baseline, we use BM25 for document retrieval using only document-level indexing information, i.e., without the benefit of the dimensional indexing model and the ability to search within a narrower context of document structure, and without the ability to include semantic evidence from concepts.

Next, we evaluate the effectiveness of using context with our dimensional data model by creating a composite retrieval score based on a linear combination of document (BM25), variable length passage of text, and sentence scores. We only use the QTM measurement for passages and sentences, i.e., no concept-based evidence.

Finally, we evaluate the effectiveness of a composite retrieval score based on a linear combination of document, passage, and sentence scores using concept-based semantic evidence. Results are measured in mean average precision (MAP).

Indexing and query processing applications were developed in Java using the Oracle 10g database. The system is platform independent, and indexes approximately 150,000 citations/hour on a 3.1GHz 2GB Pentium 4 PC.

VI. RESULTS

A. Baseline

The system delivered baseline results of 0.302 mean average precision (MAP) on the genomics collection for document retrieval using the BM25 retrieval function (k1=1.4, k3=7, b=0.75). This establishes a high-performing baseline for document retrieval which exceeds the top result from the 2005 Genomics Track at TREC [11] of 0.288 by 4.9%. We attribute much of the improvement of our baseline over the top TREC result to our acronym expansion technique. This acronym technique improved mean average precision (MAP) search results by 1.8 points (6.4% improvement) in our evaluation. The remaining improvement may be due to improved gene/protein term normalization, and BM25 retrieval function parameterization.

B. Context evaluation with dimensional data model

BM25 was used for document retrieval, and QTM was used for passage and sentence retrieval. Document, passage, and sentence retrieval scores were normalized prior to use in the linear weighted sum to determine the final composite ranking score. Table 2 shows the results from our dimensional model evaluation with several weighting strategies for combining document W_d, passage W_p, and sentence W_s scores.

TABLE 2. DIMENSIONAL MODEL RANKING

SC_d	W_d	W_p	W_s	MAP	% imp.
BM25	1	0	0	0.301	-
BM25	.75	.25	0	0.315	4.65%
BM25	.5	.5	0	0.317	5.32%
BM25	.25	.75	0	0.316	4.98%
BM25	0	1	0	0.275	-8.64%
BM25	.75	0	.25	.314	4.32%
BM25	.5	0	.5	.315	4.65%
BM25	.25	0	.75	.311	3.32%
BM25	0	0	1	.280	-6.98%
BM25	**.5**	**.25**	**.25**	**.319**	**5.98%**
BM25	1	1	1	.318	5.65%

Our best performing composite score delivered an improvement of 5.98% over our top baseline measurement and an improvement of 10.8% over the top performing result from TREC. All of top performing scores combined contextual evidence from each level of document structure, i.e., document, passage, and sentence. As long as a retrieval function included information from each level, it was relatively insensitive to specific weighting parameters.

C. Rule based ranking with semantic evidence

Finally, we evaluated our rule based ranking strategy combining concept-based semantics, context, and term statistics. The results shown in Table 3 increased MAP to 0.328 and 13.89% above the top TREC genomics track result from 2005, and 44.6% over the track average.

TABLE 3. RANKING WITH SEMANTIC EVIDENCE

Retrieval Method	MAP	% imp.
Top TREC Result	0.288	-
BM25	0.302	4.86%
Composite w/ Context	0.319	10.76%
Ranking w/ Semantics	**0.328**	**13.89%**

The average precision for each individual query showed results that statistically matched and in most cases exceeded the average precision of each query from the 2005 TREC Genomics track. As shown in Table 4, 31 queries exceeded the track average MAP for individual queries by > 10%. 74%

of these queries utilized semantic evidence from concepts identified (within the query) by the query processing algorithm. Fully 100% of queries that exceeded the track average by >50% using concept-based semantics.

TABLE 4. IMPROVEMENT OVER TRACK AVERAGE

% improvement over Track MAP	# queries exceeding	# queries exceeding w/ concepts
> 10%	31	23
> 50%	25	25
>100%	11	11

These results demonstrate the retrieval model's ability to significantly improve results with increasing amounts of semantic evidence, and perform at or above baseline with no or varying levels of semantic evidence.

VII. DISCUSSIONS AND RELATED WORK

Using retrieval of fixed length passages of text to improve retrieval of relevant documents is based on the premise that only a small portion of each relevant document is relevant to a user's query. Similarity coefficients are computed at the passage level, and the highest scoring passage or some combination of the scores of individual passages is used to compute a document's similarity coefficient [12,13,14]. Callan used a combination score with document and passage level evidence to obtain their best results [15]. These efforts focused on fixed length passages of text and did not include multiple levels of document context and semantic evidence. Tellex performed a quantitative evaluation of passage retrieval algorithms used by question-answering systems. Common to all three top performing algorithms is a non-linear boost to query terms that occur very close together in a candidate passage [16].

Mayfield and Finn advocated an approach for search on the semantic web where in the absence of semantic markup, their system would rely on traditional information retrieval techniques [17]. Regev, et al., utilized a rule-based information extraction technique for identifying gene names in text [18]. Building a search engine on top of relational technology is covered by Grossman and Frieder [19].

IX. CONCLUSION

We presented a novel information retrieval model for combining semantics, term statistics, and context for improving search precision of genomics literature. Results exceeded the state-of-the-art by 13.89% as assessed by the TREC 2005 Genomics track.

The results demonstrate improved search results in the presence of varying levels of concept-based semantic evidence, and the model still performs at or above baseline in the absence of semantic evidence. Results also show higher performance using retrieval functions that combine document as well as sentence and passage level information versus using document, sentence or passage level information alone.

The system can be efficiently implemented with a standard relational database on commodity PC hardware.

REFERENCES

[1] S. Davidson, C. Overton, P. Buneman, "Challenges in integrating biological data sources," Journal of Computational Biology, 2(4), 557-572, 2005.

[2] National Center for Biotechnology Information (NCBI), http://www.nlm.nih.gov.

[3] D. Grossman, O. Frieder, "Information Retrieval: Algorithms and Heuristics," Second Edition; Springer Publishers, ISBN 1-4020-3003-7, 1-4020-3004-5, 2004.

[4] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M Venckatrao, F Pells, "Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. Data Mining and Knowledge Discovery," Volume 1, Issue 1, 1997.

[5] R. Kimball, "The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses," Ralph, John Wiley, 1996.

[6] A. Schwartz, M. Hearst, "A simple algorithm for identifying abbreviation definitions in biomedical text," Pacific Symposium on Biocomputing, 2003.

[7] J. Urbain, N. Goharian, "A Relational Genomics Search Engine," BIOCOMP 2006: 69-74.

[8] M.F. Porter, "An algorithm for suffix stripping," Program, 14:130–137, 1980.

[9] S. Robertson, S. Walker, "Okapi/Keenbow at TREC-8," NIST Special Publication 500-246, 2000.

[10] J. Urbain, N. Goharian, O. Frieder, "IIT TREC-2006: Genomics Track," Proceedings of the Fifteenth Text REtrieval Conference, 2006.

[11] W. Hersh, et al., "TREC 2005 Genomics track overview," Proceedings of the Fourteenth Text REtrieval Conference, Gaithersburg, MD, 2005.

[12] A. Ittycheriah, S. Roukos, "IBM's Statistical Question Answering System," TREC-11, 2001.

[13] M. Kaszkiel, J. Zobel, "Passage retrieval revisited," Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2001.

[14] J. Lin, "The Role of Information Retrieval in Answering Complex Questions," Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistic, 2006.

[15] J. Callan, "Passage-Level Evidence in Document Retrieval," Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1994.

[16] S. Tellex, B. Katz, J. Lin, A. Fernandes, G. Marton, "Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering," Proceedings of the 26th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2003.

[17] J. Mayfield, T. Finn, "Information retrieval on the Semantic Web: Integrating inference and retrieval," Proceedings of the 26th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2003.

[18] Y. Regev, et al., "Rule-based Extraction of Experimental Evidence in the Biomedical Domain – the KDD Cup 2002 (Task 1)," SIGKDD Explorations, Vol. 4, Issue 2, 2002.

[19] D. Grossman, O. Frieder, "Integrating structured data and text: a relational approach," Journal of the American Society for Information Science, Volume 48 , Issue 2, February 1997.