



***IIT TREC-2007 Genomics Track:  
Using Concept-based Semantics in  
Context for Genomics Literature  
Passage Retrieval***

Jay Urbain, Nazli Goharian, Ophir Frieder

**Information Retrieval Lab**

**Illinois Institute of Technology**

{urbajay,goharian,frieder}@iit.edu



# Overview



Passage retrieval model for combining concept-based semantics and term statistics in context.

Objective: Passage retrieval precision

Methods:

- Datawarehouse style indexing for *efficient* search and aggregation of multi-word terms, or resolved concepts at multiple levels of document granularity.
- Rule-based query processing algorithm for parsing, identification, and extraction of biological concepts.
- Retrieval function for systematically combining concept terms with term statistics at multiple levels of context.
- Dependency grammar analysis for identifying complementary subject/object concept pairs.



# Outline



- Rationale
- Indexing model
- Indexing process
- Query processing
- Methods
- Results



# Rationale - Research Component

Biological research has been transformed by the explosion of scientific literature documenting the results of research facilitated by high-throughput techniques.

As a result, accurate extraction and retrieval of information from genomics text has become a key component in experiments.

Required for identifying the interplay between genes, proteins, and other biological and disease processes.



# Genomics Retrieval is still hard

- Wide variation of synonyms, acronyms, and morphological term variants for identifying similar concepts:
  - *Bovine spongiform encephalopathy, BSE, MCD, Mad Cow Disease, JCD, CJD, Creutzfeld-Jakob disease, etc.*
  - *Apolipoprotein E, ApoE, Apo-E, Apo E.*
  - *Phosphatase, PP , PhoA*



## Retrieval is hard (2)

- Acronyms can have multiple meanings (polysemy) and require contextual clues for disambiguation:
  - *IP: immunophenotype, intraperitoneally, immunoprecipitation, inositol phosphates, ischemic preconditioning, inverted papilloma.*
  - *GF: germ free, grip force, GDNF family, griseofulvin, growth factor, gel filtration, glycolytic flux, growth fraction, Granuloma faciale, gingival fibroblasts, glia filament preparation.*



# Limitations of Knowledge Sources

NCBI Databases can be helpful in providing semantic evidence for identifying named biological entities.

Unfortunately, no knowledge source can

- fully capture the complexities of language,
- accurately mimic actual term use,
- nor be up-to-date with the dynamic vocabulary of an evolving science.



# Varying levels of semantics

The availability *semantic evidence* in text can vary, making accurate identification of biological concepts difficult.

Optimal retrieval solutions require integration additional sources of evidence:

- Identification of key phrases and terms within multiple levels of context
- Leverage probabilistic measures of relevance.
- Integrate external knowledge sources

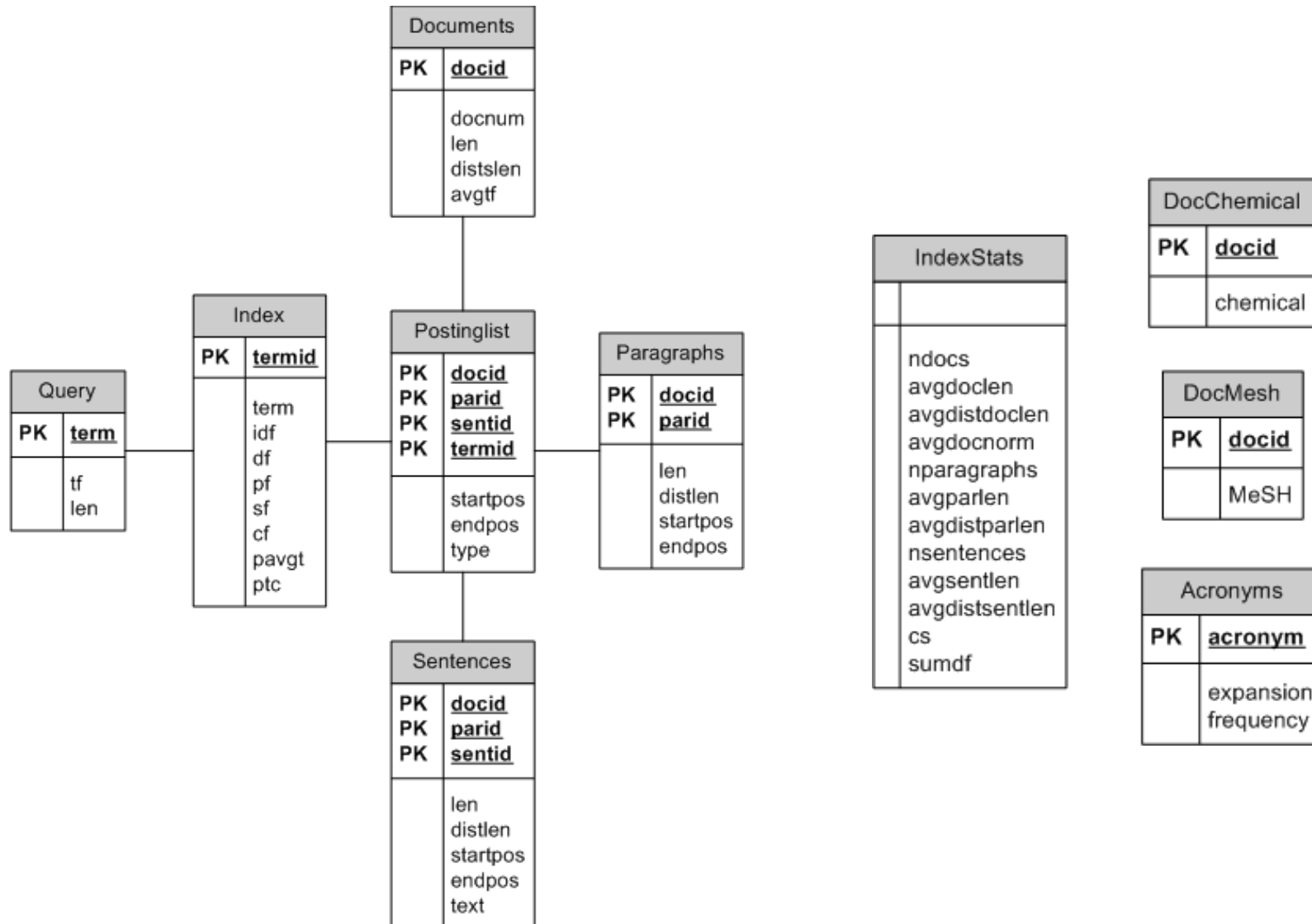


# Indexing Model

- Paragraphs, sentences, and terms represent complete topics, thoughts, & units of meaning.
- Provide logical breakdown of lexical structure into finer levels of meaning and content.
- Capture these hierarchical relationships within a search index based on a dimensional data model.



# Data Model





# Indexing Process

## *Lexical Partitioning:*

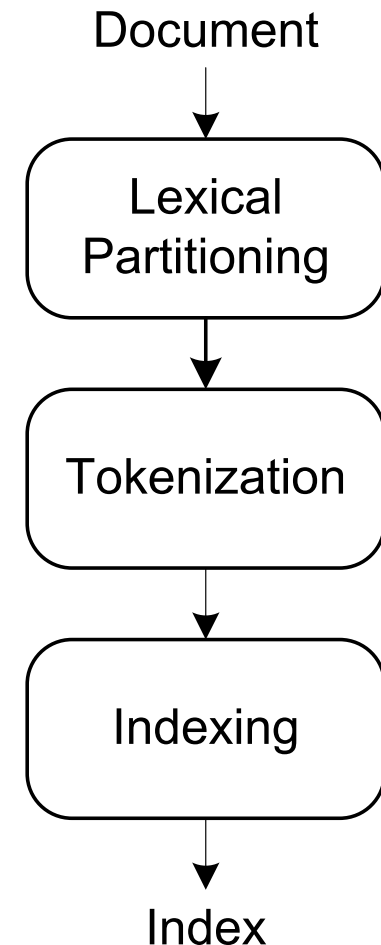
- Documents are parsed into paragraphs, & sentences.

## *Acronym Identification & Tokenization:*

- Acronyms and long-forms identified (Schwartz and Hearst): *immuno deficiency enzyme (IDE)*
- Terms tokenized, gene/protein terms normalized stop words removed, and lexical variants are generated:
  - TGF-beta1 -> tgfbeta1, tgfbeta, tgf beta1, tgf beta 1.

## *Indexing:*

- Words, long-form expansions, and lexical variants are *indexed at same position* within index.





# Query Processing

Sample query: “*Provide information about the role of the gene PRNP (prion protein) in the disease Mad Cow Disease*”.

1. Sentences are extracted, acronyms and their long-forms are identified: *PrnP (Prion Protein)*.

2. Part-of-speed tagging (HMM 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. Noun-phrase chunking  
*[gene PRNP], [prion protein], [Mad Cow Disease]*.

Natural Language Query

Sentence  
Extraction

Part of  
Speech Tag

Noun Phrase  
Chunking

Concept  
Resolution

Structured Query



# Query Processing (2)

## 4. Concept resolution

- Candidate concepts are first verified in the index
- Candidate synonyms are identified:
  - Acronym table generated during indexing
  - External sources: UMLS®, OMIM™, and Entrez Gene
  - *Synonyms considered ambiguous are not included:*
    - *Normalized IDF < 0.1.*
    - *Synonym correlates < 50% with long form.*



# Structured Concept Query



## Structured Query

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



# Keyword Search

BM25 probabilistic algorithm (Robertson and Walker, 2000) for retrieving documents, paragraphs, or sentences using keyword search.

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

- $k_1=1.4$ ,  $k_2=0$ ,  $k_3=7$ , and  $b=0.75$ .



# Paragraph Retrieval

BM25 SQL generated for document retrieval using dimensional indexing model:

```
select p.docid, p.parid, max(d.docnum) docnum,  
       sum( ln((s.nparagraphs-i.df+0.5)/(i.df+0.5))*  
            (((k1+1)*p.tf)/(k1*((1-b)+b*(par.len/s.avgparlen))+p.tf))*  
            ((k3+1)*q.tf/(k3+q.tf)) ) as sc  
from index i, postinglist p, documents d, paragraphs par, query q  
where p.docid=par.docid  
and p.docid=par.parid  
and p.docid=par.docid  
and i.termid=p.termid  
and i.term=q.term  
group by p.docid, p.parid  
order by sc desc;
```



# Concept Extraction

Sample query:

*Exact reactions that take place when you do glutathione S-transferase (GST) cleavage during affinity chromatography*

Concepts and term variants (stemmed form) are identified

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



# Concept Search

2) Index is searched for all term variants of *each* concept. E.g. “*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  
UNION
```

...

```
where i1.term='affin' and i2.term= 'purif '
```

...

3) Index resolved concepts in *conceptlist*.



# Passage Identification

- 1) Min-spanning tree constructed from the max number of *distinct* concepts within the shortest lexical distance:

*affinity chromatography, and purified Mce1A and Mce1E, free of the fusion partner, were recovered following specific proteolytic cleavage of the **GST***

- 2) Passages 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.*



# Passage and Sentence Ranking

- Distinct number of concepts identified within the passage or top sentence.
- Sum of normalized IDF's of concept terms.
- *Query term density match (QTM)* measurement we devised for TREC 2006.

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

- Sums the normalized IDF's of each *distinct* matching *query* term or concept term at the sentence level.
- Passage similarity coefficients (SC) are aggregated as the top 3 sentence scores.



# Identifying Dependent Concepts

- We used Stanford's<sup>1</sup> dependency grammar parser to identify subject/object complements between queries and passage sentences.
- Dependencies are motivated by grammatical function, i.e., syntactically and semantically. A word depends on another if it is either a complement or a modifier of the latter.
- If we can identify the modifier of the object in a passage sentence that corresponds to the subject in the original query, we can (*hopefully*) increase the likelihood of answering the query.

<sup>1</sup>Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. *Generating Typed Dependency Parses from Phrase Structure Parses*. In LREC 2006.



## Dependent Concepts Example

Example, query 201: *What [mutations] in the **Raf gene** are associated with **cancer**?*

Retrieve the following passage MST:

*...**melanoma** cell lines with **B-RAF** and **N-RAS** mutations...*

for which we can identify dependencies between the modifiers **B-RAF** and **N-RAS** and the object of the sentence **mutations** which was the subject of the original query.



# Composite Ranking

- First, we applied the *dependency grammar passage rank boost*:
  - Passages containing sentences where we have identified *query subject/object complements* are ranked above all other passages with the same number of distinct concepts.
- Second, passages are ranked using a linear normalized weighting of *distinct number of concepts*, *summation of distinct concept NIDF*, or *QTM*.

$$SC_{composite} = w_1 SC_1 + w_2 SC_2 + \dots + w_n SC_n$$

- Finally, passages with the same  $SC_{composite}$  are ordered by a passages' lexical distance, i.e., the width of the MST of distinct concept instances.



## Submitted Runs

Run	Retrieval Function
<b>iitx1</b> <i>emphasize passage</i>	0.45*distinct number of passage concepts + 0.05*distinct number of sentence concepts + 0.45*passage norm IDF sum + 0.05*sentence norm IDF sum <i>no dependency grammar passage rank boost</i>
<b>iitx2</b> <i>emphasize top sentence with dependency grammar boost</i>	0.05*distinct number of passage concepts + 0.45*distinct number of sentence concepts + 0.05*passage norm IDF sum + 0.45*sentence norm IDF sum <i>dependency grammar passage rank boost</i>
<b>iitx3</b> <i>emphasize passage with dependency grammar boost</i>	0.45*distinct number of passage concepts + 0.05*distinct number of sentence concepts + 0.45*passage norm IDF sum + 0.05*sentence norm IDF sum <i>dependency grammar passage rank boost</i>



## Official Results for runs submitted to TREC (% above track median)

<b>Run</b>	<b>Document MAP</b>	<b>Aspect MAP</b>	<b>Passage MAP</b>	<b>Passage2 MAP</b>
iitx1	0.2454 (31.15%)	0.1272 (18.06%)	0.0852 (75.27%)	0.0388 (39.82%)
iitx2	0.2462 (31.60%)	0.1166 (8.16%)	0.0926 (90.38%)	0.0335 (20.56%)
iitx3	0.2414 (28.99%)	0.1253 (16.25%)	0.0940 (93.22%)	0.0443 (59.30%)




## Corrected results for run IITx3

We discovered an error in our database software where frequently occurring terms were not stored in our index. Such terms included *gene* and *protein*. The corrected results for *IITx3*

<b>Corrected Run</b>	<b>Document MAP</b>	<b>Aspect MAP</b>	<b>Passage MAP</b>	<b>Passage2 MAP</b>
iitx3	0.2670	0.1662	0.1060	0.0616



# Postmortem

- 1) Underperformance for queries where we did not identify important concepts.
  - For example, query 235: *Which [genes] involved in NFkappaB signaling regulate iNOS?*
  - We identified only 81/182 relevant passages largely due to not identifying variations of **NF-kappaB**:
    - **NF-** **B**
    - **NF-kB**



# Postmortem

2) Underperformance due to ranking, i.e., we identified the relevant passages, but did not rank them well.

- For example, query 209: *What [biological substances] have been used to measure toxicity in response to etidronate?*
- We identified *72 out of 78* relevant passages, but we still underperformed, **but....**
  - Our heavy weighting of distinct number of concepts ranked passages with more general concepts like “biological substances” and “toxicity” versus “etidronate” too high.
  - Subsequent runs with heavier weighting of *concept IDF* and *QTM* improved document and passage MAP.



# Postmortem

- 3) Our MST passage identification algorithm, where passages were expanded to sentence boundaries appeared to work well.
- Many relevant passages were identified exactly.
  - The algorithm had trouble in the following cases:
    1. Where a relevant passage was identified as a large block of references.
    2. Where relevant passages were identified as less than a sentence – not many, and these were hard to understand even for a human.



# Postmortem

- 4) Dependency grammar boosting increased performance by 10.3% (*iitx3* versus *iitx1*).
- This result was statistically significant using a *paired-t test*, however the passage retrieval precision (9-10%) was relatively low...
  - ...and with passage retrieval precision ~10%, there's a lot of room for improvement.
  - Technique deserves further evaluation.



# Conclusion & Future Efforts

- Solid results:
  - Still need better concept resolution.
  - Higher weighting of query term statistics, and “more” important concepts.
  - Good passage identification algorithm.
  - Dependency grammar shows promise.
  - Dimensional data model efficient, flexible for generating retrieval functions.

## *Future Efforts*

- Develop more formal probabilistic topic and concept models.
- Incorporate additional types & sources of evidence



# Collection

- 162,259 documents from 49 biology journals published by Highwire Press (HTML format).
- 12,641,127 maximum legal spans corresponding to paragraph boundaries.
- 36 topics from practicing biologists (culled from 50 collected).
- Queries phrased as list entity-based questions.
- Relevance judgements performed by 13 biology PhD students.



# Related Work

- Using retrieval 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.
- Callan used a combination score with document and passage level evidence to obtain their best results.
- Tellex performed a quantitative evaluation of passage retrieval algorithms used by question-answering systems. All three top performing algorithms used a non-linear boost to query terms that occur very close together in a candidate passage.



## Related Work (2)

- 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.
- Regev, et al., utilized a rule-based information extraction technique for identifying gene names in text.
- Building a search engine on top of relational technology is covered by Grossman and Frieder.



# References

*Callan, James P., 1994. Passage-Level Evidence in Document Retrieval, Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

*Grossman, D., Frieder, O., Holmes, D., Roberts, D., 1997. Integrating Structured Data and Text: A Relational Approach, JASIS, 48(2).*

*Hersh, et al. (2005). TREC-2005 Genomics Track Protocol.*

*Ittycheriah, Abraham, Roukos, Salim, 2001. IBM's Statistical Question Answering System, TREC-11.*

*Kaszkiel, M., Zobel, J., 1997. Passage retrieval revisited, Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval.*

*Kaszkiel, Marcin, Zobel, Justin, 2001. Effective Ranking with Arbitrary Passages, Journal of the American Society of Information Science.*

*Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In LREC 2006.*

*Porter, M.F., 1980. An algorithm for suffix stripping, Program, 14:130–137.*

*Robertson S.E., Walker, S., 2000. Okapi/Keenbow at TREC-8, NIST Special Publication 500-246.*

*Schwartz, Ariel, Hearst, Marti. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text, Pacific Symposium on Biocomputing, Kauai.*



# Supporting Evidence - Passages

- Effectiveness of passage and aspect retrieval is largely dependent upon initial document retrieval stage.
- Using passage-based retrieval 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.
- Several techniques have been used to define passages – all focused on defining a more narrow context (Callan, 1994; Hearst, 1994; Ittycheriah, et al., 2001; Kaszkiel and Zobel, 2001; Kaszkiel and Zobel, 1997; Lin, 2006; Tellex, et al., 2003; White, et al., 2005).
- Passage retrieval has also been shown to be a key step for identifying the proper context for question-answering systems