Knowledge-intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature

Wei Zhou, Clement Yu
Neil Smalheiser, Vetle Torvik, Jie Hong

University of Illinois at Chicago

Motivation

To answer biologists' information needs which are expressed as complex questions.

Start with:

"What is the role of gene PRNP and Mad Cow Diseases?"

Look for:

"… PRNP is an agent causing fatal neurodegenerative disorders such as mad cow diseases (MCD) …"

PubMed
A service of the National Library of Medicine and the National Institutes of Health
www.pubmed.gov

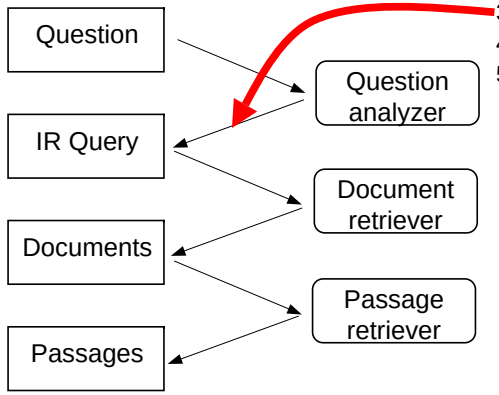
1. Query understanding is limited
2. No relevance ranking
3. No passage extraction

Google

- No query understanding

… The prp gene encodes the putative causative agent of the transmissible spongiform encephalopathies (tses), a heterogeneous group of fatal, neurodegenerative disorders including human Creutzfeldt CJakob disease, bovine spongiform encephalopathy, ovine scrapie and chronic wasting disease (CWD) of North American deer and elk. …

Approach



Query expansion
1. Synonyms
2. Hyponyms
3. Hypernyms
4. Lexical variants
5. Implicitly related terms

Why is it important to recognize concepts and expand them with their alternatives or related terms ?

- 1/3 of term occurrences are variants (Jacquemin et al., 2001)
- Having 6-7 synonyms for a single gene is not unusual in the domain of Genomics (Bernardi et al., 2002)
- The probability of two experts using the same term to refer to the same concept is less than 20% (Grefenstette et al., 1994)

Methodology

1. Query concepts identification
2. Query formulation (or expansion)
3. Conceptual IR model
4. Passage extraction

Definitions

1. A **concept** is an entry term in the biomedical thesauruses. (MeSH and Entrez gene)

2. “Mad Cow Disease” has a **semantic type** of “Disease or Syndrome”.

Query concepts identification

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

Templates have been used for identification of genes.

Query expansion

- Synonyms
- Lexical variants
- Hyponyms
- Hypernyms
- Implicitly related terms

Related works on query expansion to improve biomedical IR systems

- Aronson, 1997
- Hersh, 2000
- Genomics track (2004,2005,2006)
- Demner-Fushman, 2007

Lexical variants

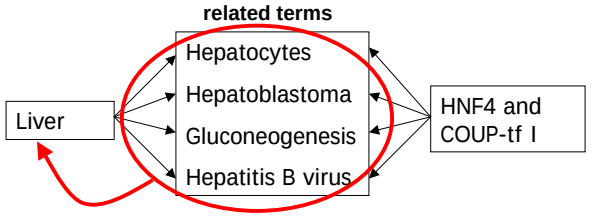
- 2 strategies:
  - Strategy 1: automatically generate lexical variants (Buttcher et al., 2004; Huang et al., 2005).
    e.g., PLA2 → PLA 2, PLAII, and PLA II
  - Strategy 2: retrieve **additional** lexical variants from a term database of MEDLINE (Zhou et al., 2006).
    e.g., PLA2 → PL-A2

Note: PLA2: Phospholipase A2

Implicitly related terms

- Terms that Co-occur frequently & related semantically

Query: How do interactions between **HNF4** and **COUP-TF1** suppress **liver** function"



Conceptual IR model

$$sim(q,d) = [sim_{concept}(q,d), sim_{word}(q,d)]$$

sim(q,d2)>sim(q,d1) if either 1 or 2: (Liu 2004)

$$1.\ sim_{concept}(q,d2) > sim_{concept}(q,d1)$$

$$2.\ sim_{concept}(q,d2) = sim_{concept}(q,d1)\ \text{AND}\ sim_{word}(q,d2) > sim_{word}(q,d1)$$

q: "role of gene **PRNP** in **mad cow disease**."

d1:

…mad cow disease ……
……mad cow disease ……
……mad cow disease ……
……mad cow disease …..
……mad cow disease…

d2:

…PRNP ……
……
……
……
……Mad cow disease……..
……
……
……
……
……
……

Okapi (likely): sim(q,d1)>sim(q,d2), but intuitively d2 is more relevant than d1.

q: “role of gene **PRNP** in **mad cow disease**.”

d1:

…PRNP ……
……
……
……
……Mad cow disease……
……
……
……
……
……
……

d2:

…PRNP ……
……
……role……PRNP……
……
……mad cow disease……
……
……PRNP……
…… mad cow disease……
……
……

Determine the weights assigned to the added concepts

$$Weight\_added\_concept = \beta * Weight\_concept$$

- β = 1 if t is the a synonym, hyponym, or lexical variant;
- β = 0.95 if t is a hypernym;
- β = 0.90*(k–i+1)/k if t is an implicitly related concept

Passage extraction

Two-step Rational:

(**Identify candidates**) Given various windows of different sizes, choose the ones which have the **maximum** number of query concepts and the **smallest** number of sentences.

(**Merging**) Merge two candidate windows if they are exactly adjacent to each other.

Results on TREC

- Query collection: 28 questions collected from biologists in 2006.
- Document collection: 162,259 Highwire full-text documents in HTML format.
- Performance Metrics
  - Passage MAP
  - Aspect MAP
  - Document MAP

Table 3.2.1 Basic conceptual IR model vs. term-based model

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

Table 3.2.2 Contribution of different types of domain-specific knowledge

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

Table 3.2.3 Contribution of abbreviation correction and pseudo-feedback

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

Results

- **1**: The basic conceptual IR model significantly outperforms the Okapi on all three levels, which suggests that, although it requires additional efforts to identify concepts, retrieval on the concept level can achieve substantial improvements over purely term-based retrieval model.

Results

- **2**: The biggest improvement comes from the lexical variants, which is consistent with the result reported in [Buttcher 2004]. This result also indicates that biologists are likely to use different variants of the same concept according to their own writing preferences and these variants might not be collected in the existing biomedical thesauruses.

Results

- **3**: The overall performance is an accumulative result of adding different types of domain-specific knowledge and it is better than any individual addition. It is clearly shown that the performance is significantly improved (107% on passage level, 63.1% on aspect level, and 49.6% on document level) when the domain-specific knowledge is appropriately incorporated.

Results

- **4**: different types of domain-specific knowledge affect different subsets of queries. More specifically, each of these types (with the exception of “the lexical variants” which affects a large number of queries) affects only a few queries. But for those affected queries, their improvement is significant. As a consequence, the accumulative improvement is very significant

Summary

- We proposed a conceptual approach to utilize domain-specific knowledge in an IR system to improve its effectiveness in retrieving biomedical literature.
- We examined the effects of utilizing concepts and of different types of domain-specific knowledge in performance contribution.

Future work

- Create a database of lexical variants
- Develop a method to recognize genes from a query.
- Systematically optimize the weighting parameter β for different types of expanded concepts.

Thank you!