Document Retrieval using Predication Similarity

Kalpa Gunaratna
Dr. Olivier Bodenreider

End of Summer Internship Presentation
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine
8600 Rockville Pike, Bethesda, MD 20894
08.07.2014
Predications and documents

• Predications are extracted from PubMed articles. They are in the triple format having a subject, a predicate, and an object.

PMID: 9688090
Pure endotoxin does not pass across the intestinal epithelium in vitro.

Entire intestinal epithelium part of Rattus norvegicus
Ileal mucous membrane part of Rattus norvegicus
Bacteria location of Lipopolysaccharides
Bacterial Translocation process of Animals
Endotoxins causes Systematic inflammatory Response Syndrome

Predication set

SemRep extracts
Motivation

• We have a repository of predications extracted by SemRep for each PubMed article. This is a knowledgebase built on top of MEDLINE.

• These predications can be used to retrieve documents using predication-predication similarity.

• This is different from document retrieval using bag of words as we are using set of predications instead.

• Objective: retrieve related documents for a given document.
Our hypothesis ...
Sim(C1,C2) - % shared ancestors
Concept-Concept similarity

- Simple measure used to compute similarity between two sets of concepts.

\[ \text{Sim}(c_1, c_2) = \frac{\text{number of shared concepts between } c_1 \text{ and } c_2}{\text{total number of concepts in } c_1 \text{ and } c_2} \]

- \( \text{Sim}(CVD, KD) = \frac{1}{3} = 0.3 \)
- \( \text{Sim}(HVD, CAD) = \frac{3}{5} = 0.6 \)
Sim(C1,C2) - % of shared ancestors

Sim(P1,P2) - average pairwise similarity of subject, predicate, and object
Predication-Predication similarity

\[
\text{Sim}(P1,P2) = Ws \times \text{Sim}(C1,C2) + Wp \times \text{Sim}(R1,R2) + Wo \times \text{Sim}(O1,O2) / (Ws + Wp + Wo)
\]

when, \( Ws = Wp = Wo = 1 \)

\[
\text{Similarity} = (0.5621 + 1 + 0.7068) / 3 = 0.7563
\]
Sim(C1,C2) - % of shared ancestors

Sim(P1,P2) - average pairwise similarity of subject, predicate, and object

Sim(Set1, Set2) - average of maximum pairwise predication similarity
Predication Set – Predication Set similarity

Similarity \( (S, D) = \frac{\sum \max_{\text{sim}}(P_S, PD) + \sum \max_{\text{sim}}(PD, P_S)}{N_S + ND} \)

\[
\text{Similarity} = \frac{(0.7 + 0.9 + 0.6 + 0.9 + 0.7)}{(2 + 3)}
\]

\[= 0.76\]
Document D1

Concept C1

Predication P1

Predication Set Set1

Document D2

Concept C2

Predication P2

Predication Set Set2

Concept – Concept & Rel – Rel similarity

Predication – Predication similarity

Predication Set – Predication Set similarity

Document – Document similarity

Sim(C1, C2) - % of shared ancestors

Sim(P1, P2) - average pairwise similarity of subject, predicate, and object

Sim(Set1, Set2) - average of maximum pairwise predication similarity

Sim(D1, D2) – predication set similarity of the two documents
PMID: 9688090
Pure endotoxin does not pass across the intestinal epithelium in vitro.

PMID: 15024697
Cholinergic, nitrergic and peptidergic (Substance P- and CGRP-utilizing) innervation of the horse intestine. A histochemical and immunohistochemical study.

Predication set

Entire intestinal epithelium
Ileal mucous membrane
Bacteria
Bacterial Translocation
Endotoxins

part of
part of
location of
process of
causes
Rattus norvegicus
Rattus norvegicus
Lipopolysaccharides
Animals
Systematic inflammatory Response Syndrome

Predication set

Large intestine
Nerve Fibers
Tissue fiber
Cell body neuron

part of
part of
part of
location of
Equus caballus (horse)
Submucous plexus
Mucous membrane
Nitric oxide
Principle of our evaluation

Our retrieval of related documents

PubMed related citation documents

comparison

30 documents

~ 900 documents

30 PubMed related citations

30 PubMed related citations

30 PubMed related citations

30 PubMed related citations
Preliminary evaluation

<table>
<thead>
<tr>
<th># retrieved documents</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>top 5</td>
<td>0.7733</td>
<td>0.1289</td>
<td>0.221</td>
</tr>
<tr>
<td>top 10</td>
<td>0.7</td>
<td>0.2333</td>
<td>0.35</td>
</tr>
<tr>
<td>top 15</td>
<td>0.6244</td>
<td>0.3122</td>
<td>0.4163</td>
</tr>
<tr>
<td>top 20</td>
<td>0.5449</td>
<td>0.3633</td>
<td>0.436</td>
</tr>
<tr>
<td>top 25</td>
<td>0.4827</td>
<td>0.4022</td>
<td>0.4388</td>
</tr>
<tr>
<td>top 30</td>
<td>0.4433</td>
<td>0.4433</td>
<td>0.4433</td>
</tr>
</tbody>
</table>
Discussion- advantages

• Semantically aware document similarity.
  • Documents as bags of predications.

• Bag of predications vs bag of words.
  • More precise.
    • “ASPIRIN TREATS HEADACHE” Vs “ASPIRIN + HEADACHE”
  • More flexible
    • “Entire intestinal epithelium” ~ “Large Intestine”

• Predication-Predication similarity as a by-product.
  • Question answering and exploration capabilities on the predication level – factual information.
    • E.g., “give me related predications to ASPIRIN TREATS HEADACHE”, “find ? TREATS HEADACHE”.
Discussion – limitations & future work

• Limitations with SemRep
  • Limited template based extraction.
  • Extracts within sentence predications.

• Limitations with similarity
  • Concept-Concept similarity needs to be tested in UMLS.
  • Predication-Predication similarity needs to be calibrated with weights.
  • More robust evaluation needed.
    • Larger and independent test collection.

• Technical limitations
  • Scaling to the whole MEDLINE and UMLS concepts.
    • Using parallel processing for computation and storage.
Acknowledgements

• I would like to thank,

  • Dr. Bodenreider for guiding the project and providing insightful advice.
  
  • Dr. Rindflesch for related research on SemRep.
  
  • Dr. Winnenburg for various technical support.
  
  • Dr. McDonald – Director, Lister Hill National Center for Biomedical Communications and Dr. Fontelo - Director, LHNCBC Training Program for supporting the internship program.
  
  • My fellow interning student Nathan Bahr for valuable comments.
Thank You

Questions?