Document Retrieval using Predication Similarity

Kalpa Gunaratna

Dr. Olivier Bodenreider

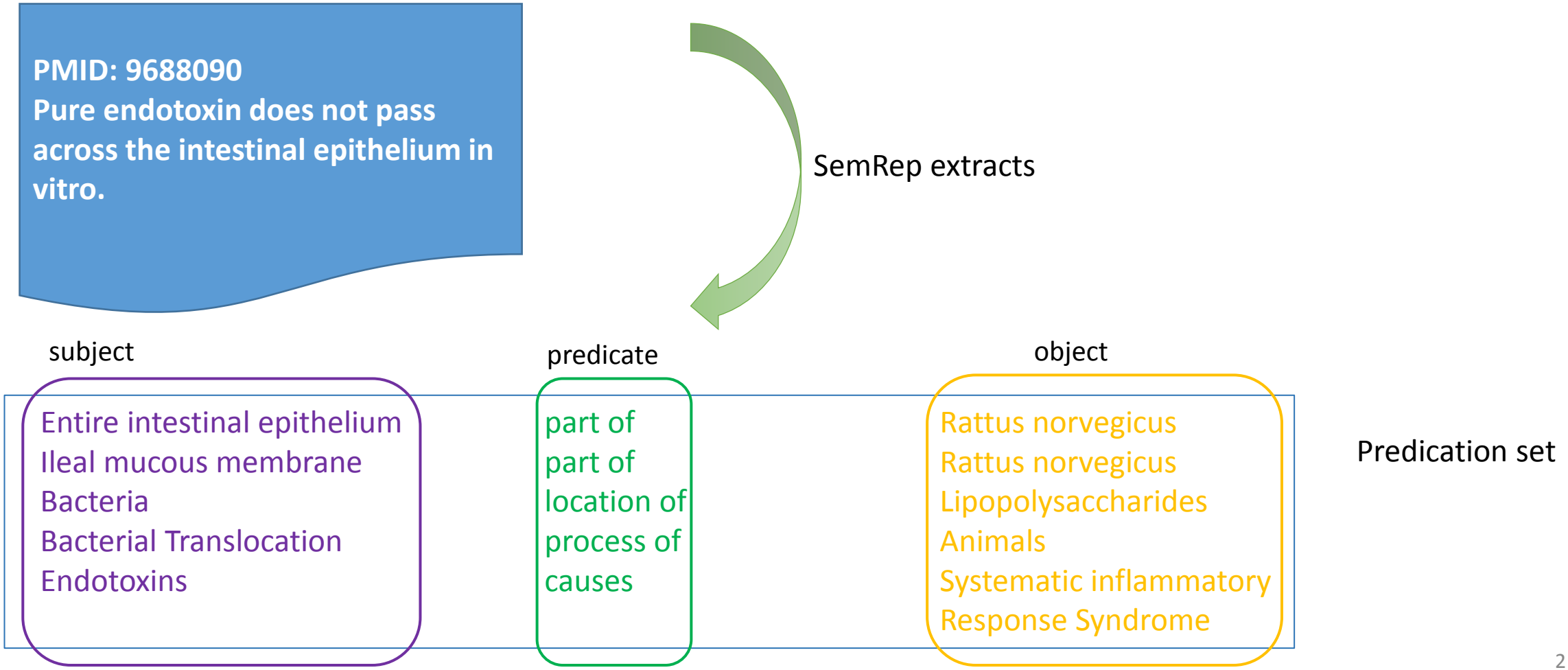
End of Summer Internship Presentation

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.



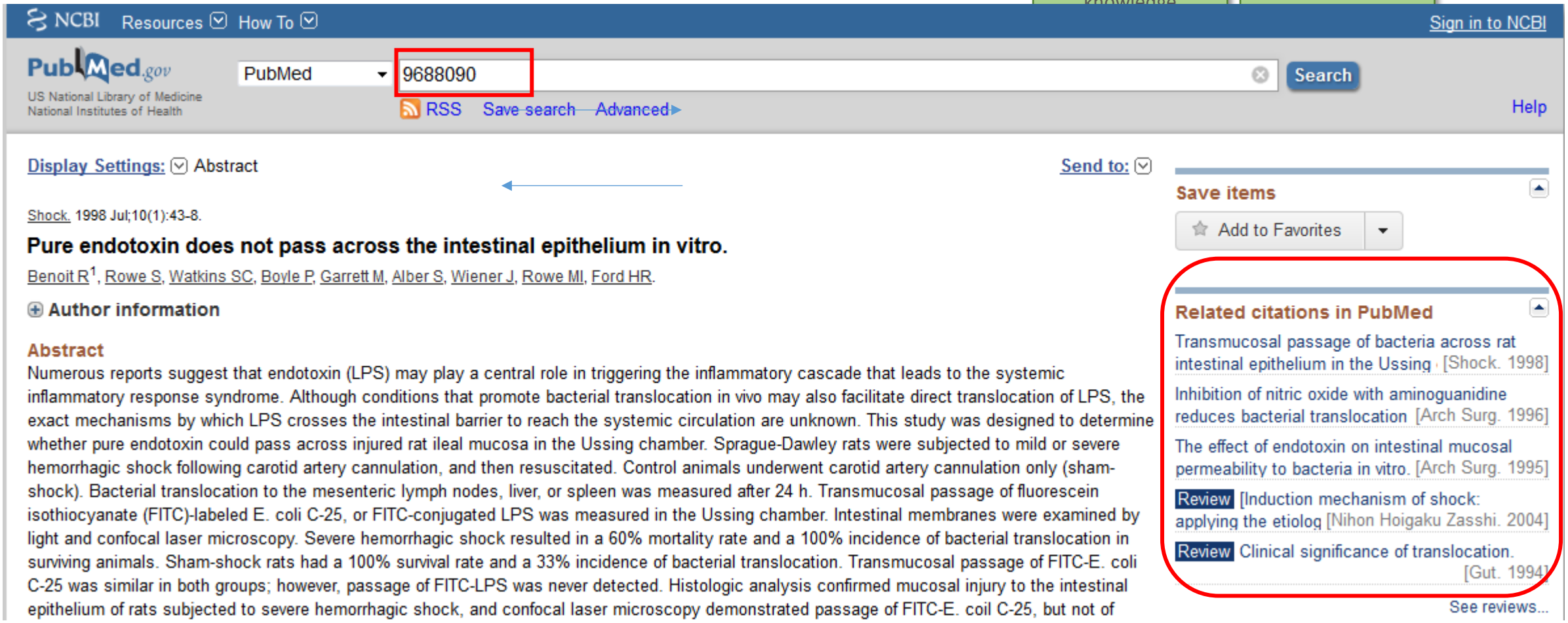
Motivation

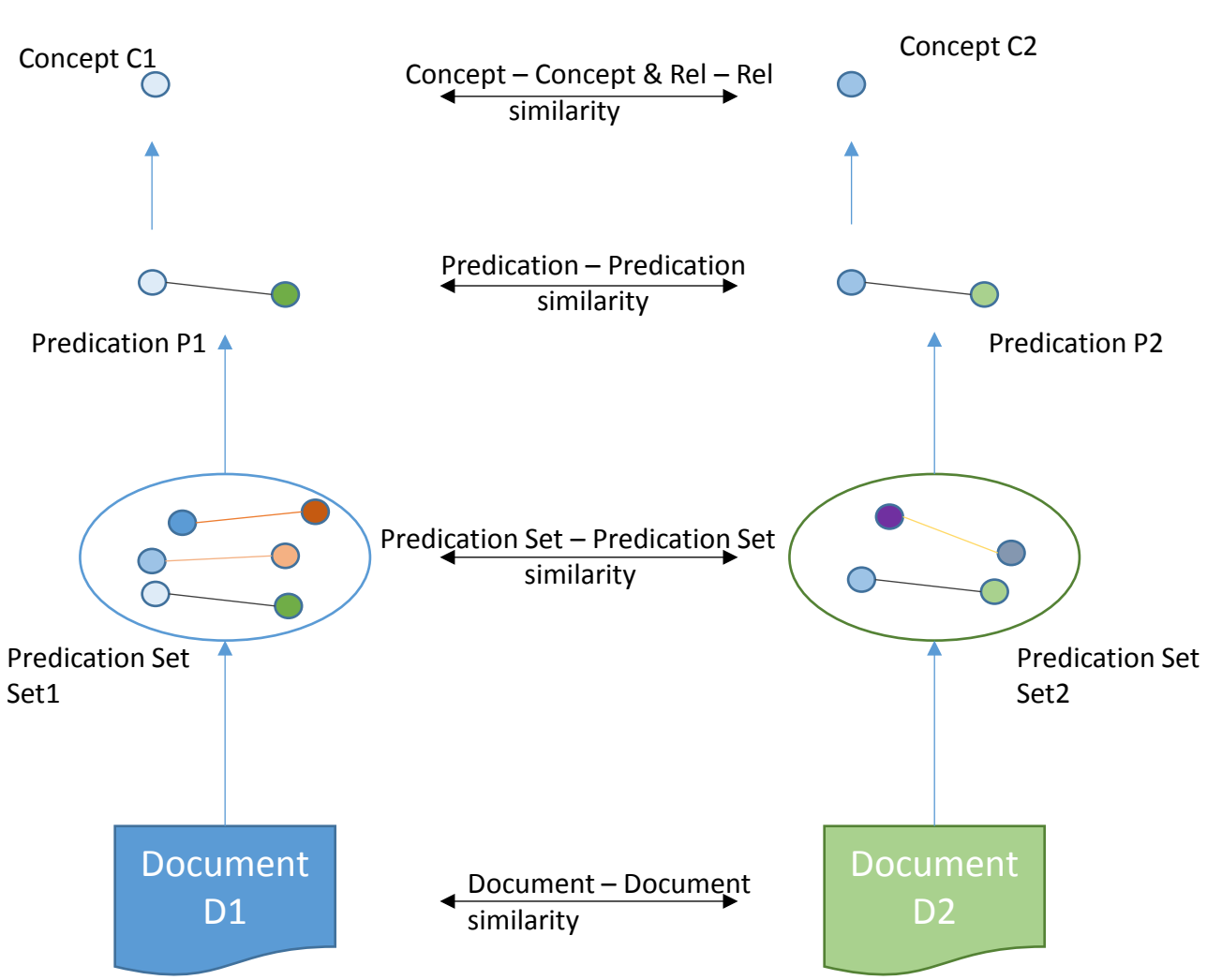
- We have a repository of predications extracted by SemRep for each PubMed article. This is a knowledgebase built on top of MEDLINE.
- These predications can be used to retrieve documents using predication-predication similarity.
- This is different from document retrieval using bag of words as we are using set of predications instead.
- Objective: retrieve related documents for a given document.

Our hypothesis …

Utilizes factual knowledge

Utilizes hierarchical

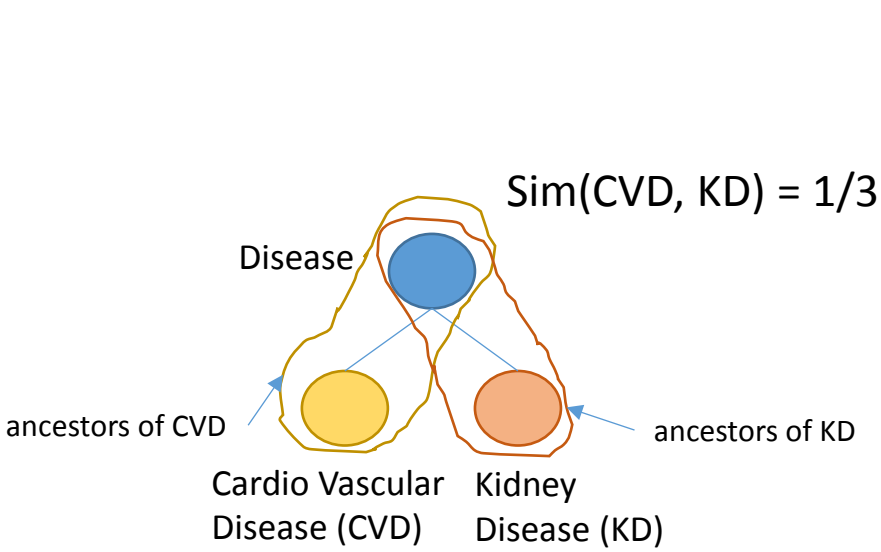


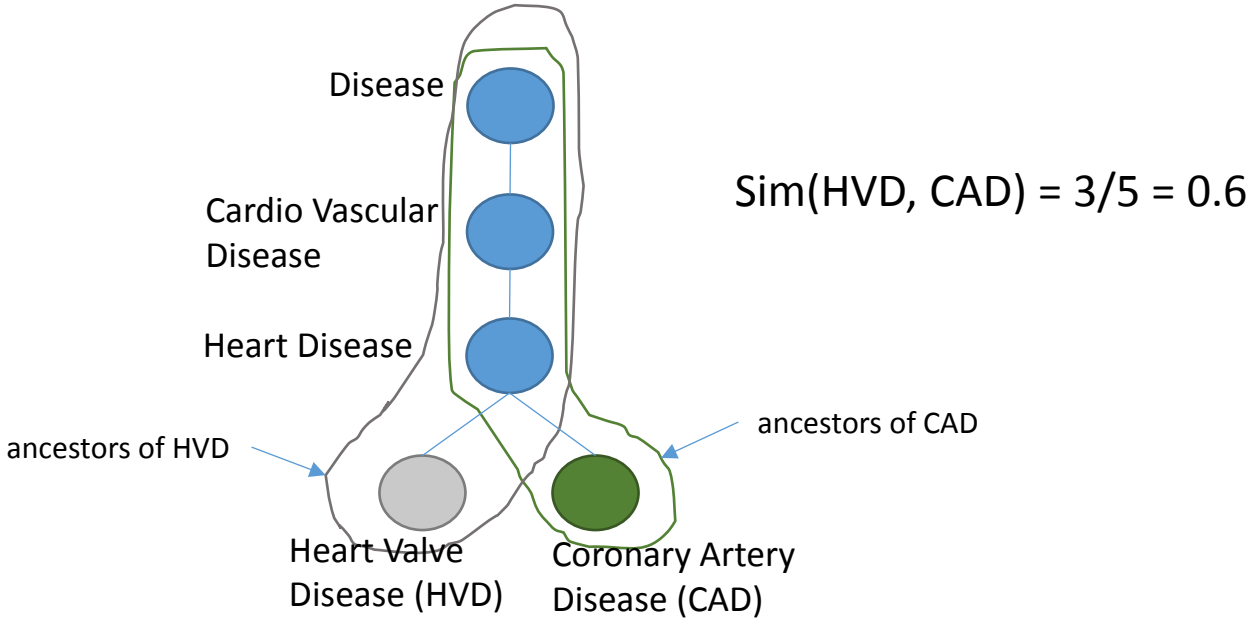


Sim(C1,C2) - % shared ancestors

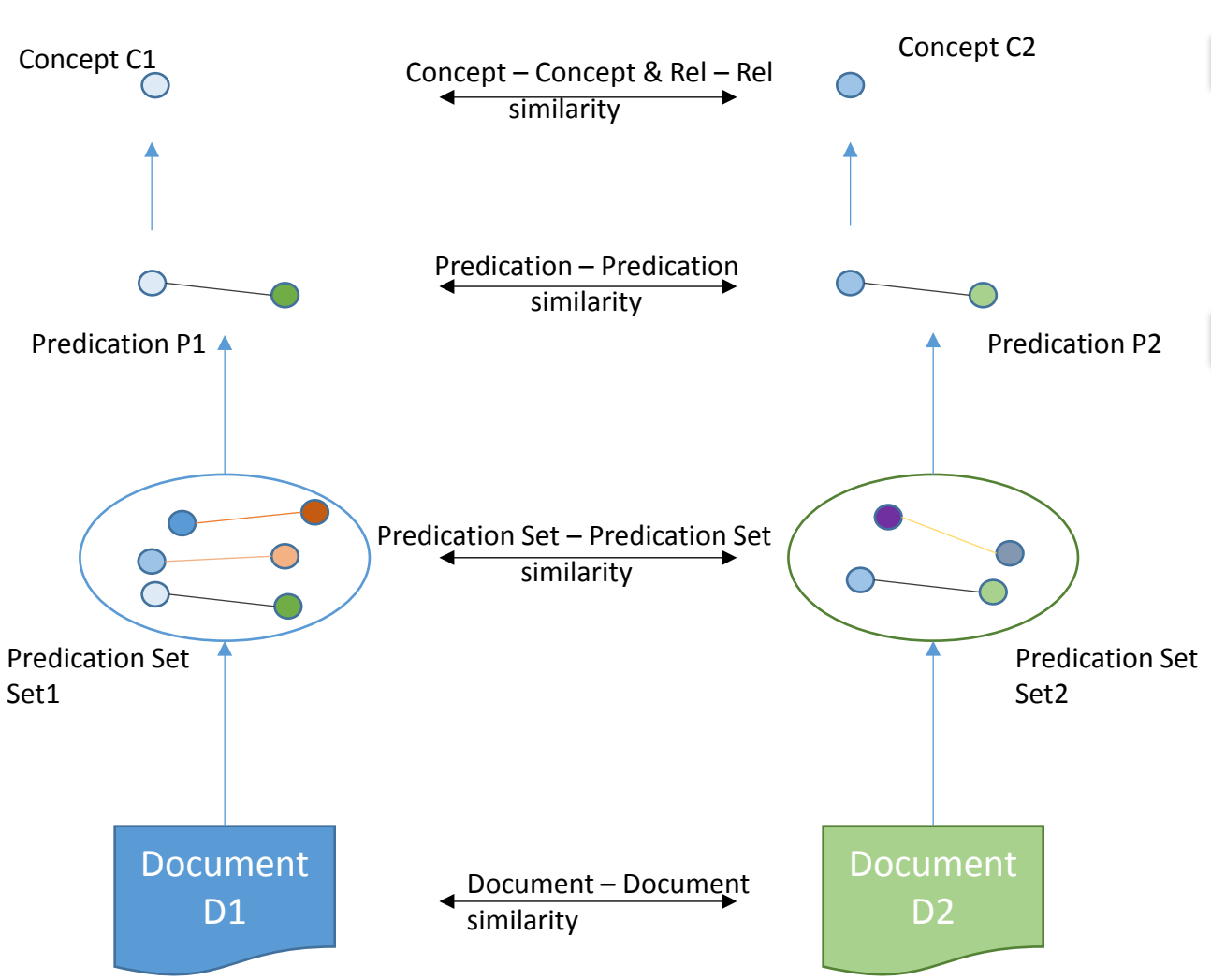
Concept-Concept similarity

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





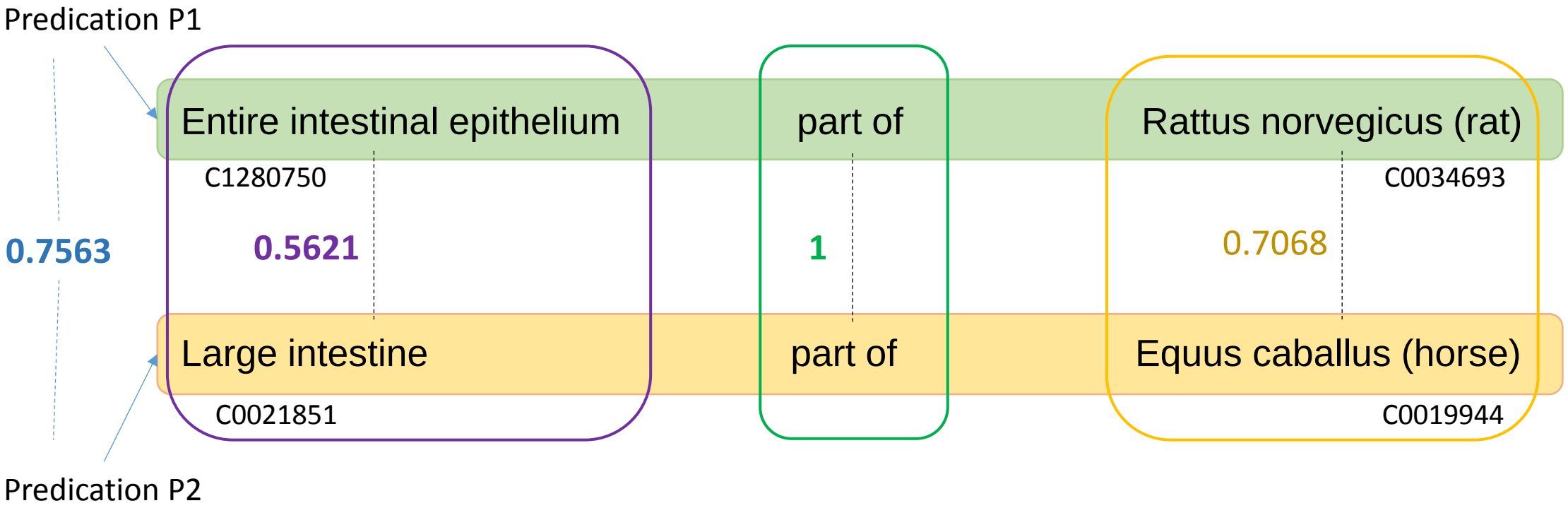
- Jaccard (c1, c2) = $\frac{number\ of\ shared\ concepts\ between\ c1\ and\ c2}{total\ number\ of\ concepts\ in\ c1\ and\ c2}$



Sim(C1,C2) - % of shared ancestors

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

Predication-Predication similarity

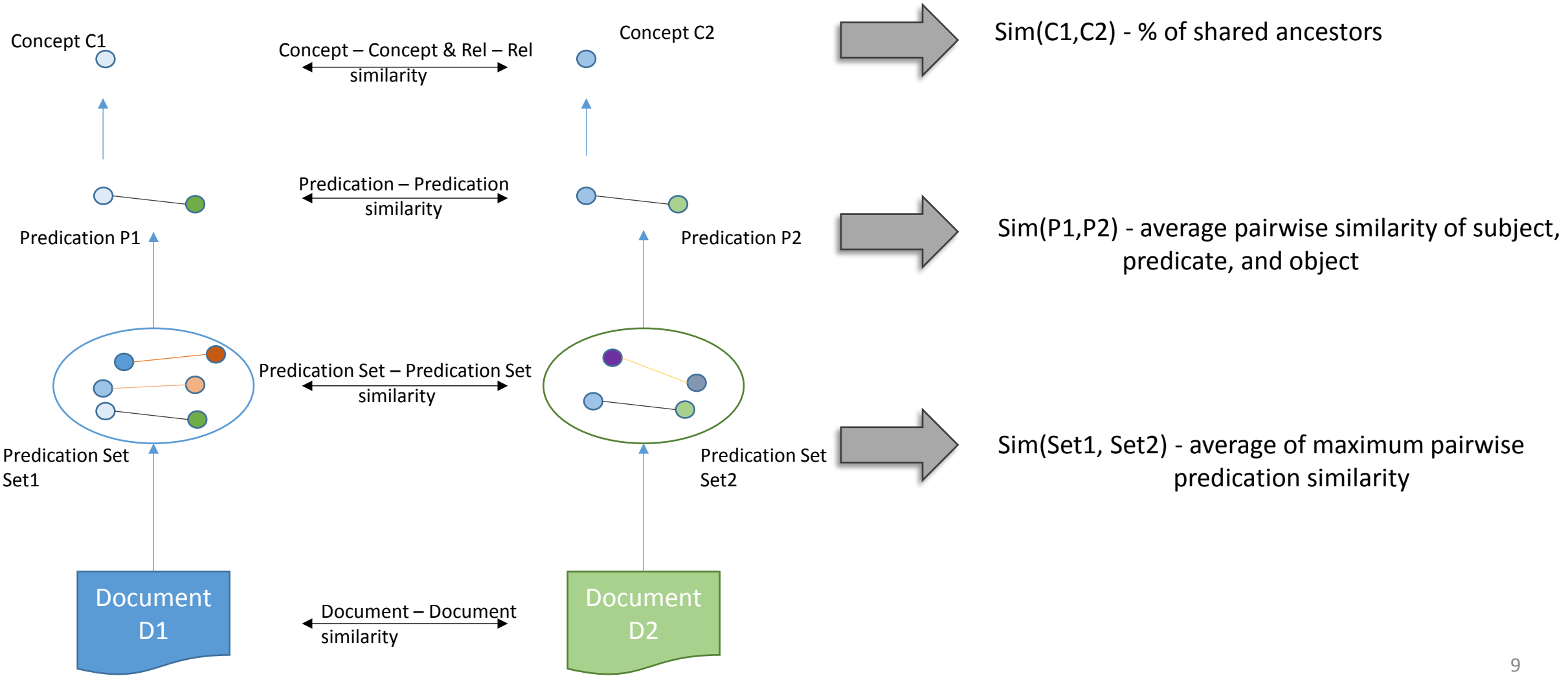


Sim(P1,P2) = Ws * Sim(C1,C2) + Wp * Sim(R1,R2) + Wo * Sim(O1,O2) / (Ws + Wp + Wo)

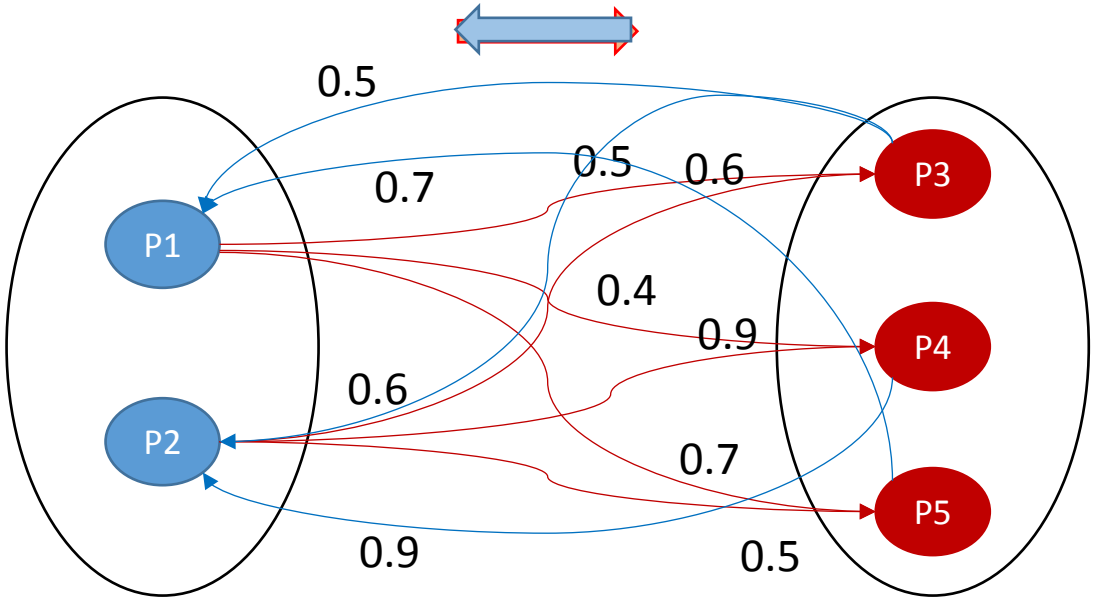
when, Ws = Wp = Wo = 1

Similarity = (0.5621 + 1 + 0.7068) / 3

= 0.7563


Concept C1
Concept – Concept & Rel – Rel similarity
Concept C2
Sim(C1,C2) - % of shared ancestors
Predication – Predication similarity
Predication P1
Predication P2
Sim(P1,P2) - average pairwise similarity of subject, predicate, and object
Predication Set – Predication Set similarity
Predication Set Set1
Predication Set Set2
Sim(Set1, Set2) - average of maximum pairwise predication similarity
Document D1
Document – Document similarity
Document D2

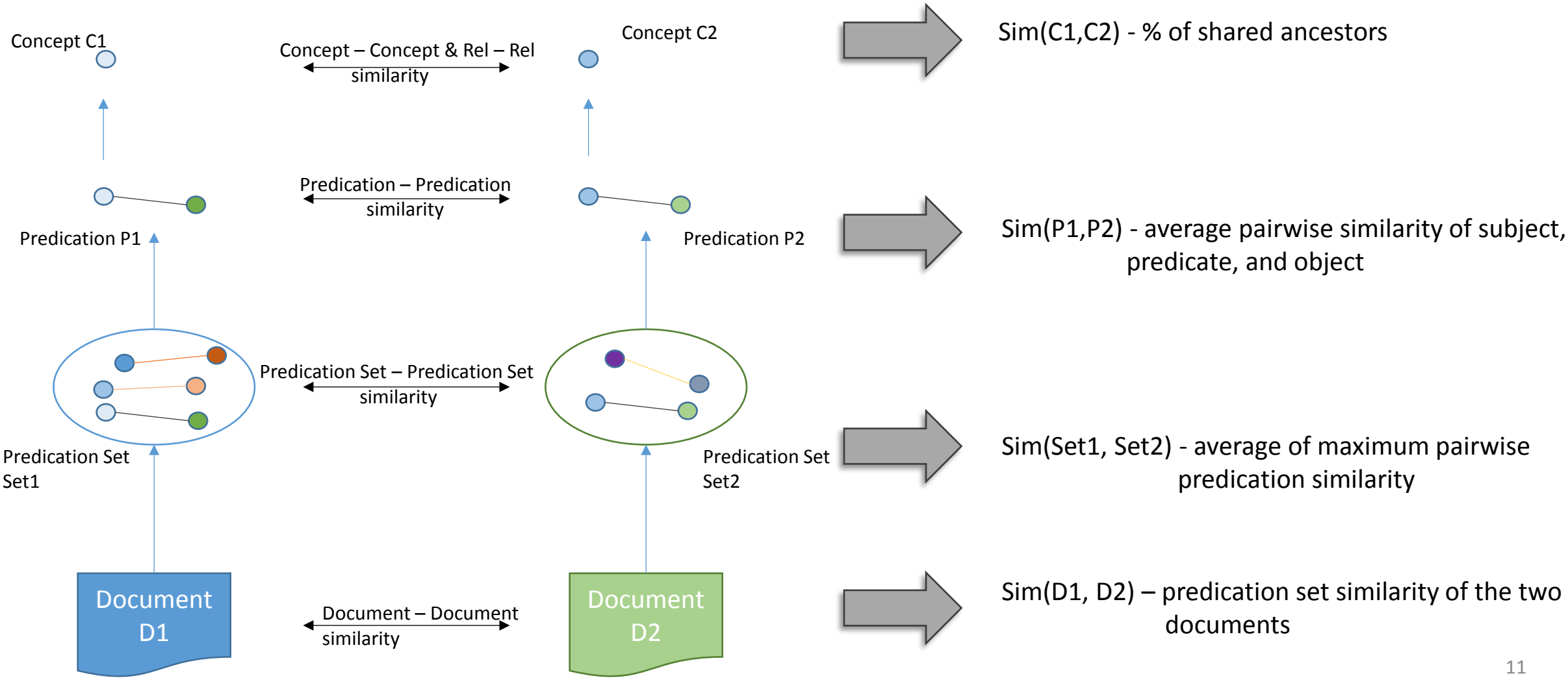
Predication Set – Predication Set similarity



Similarity = (0.7 + 0.9 + 0.6 + 0.9 + 0.7) / (2 + 3)

= 0.76

$$\text{Similarity (S, D)} = \frac{\sum \text{max\_sim}(P_S, PD) + \sum \text{max\_sim}(PD, PS)}{N_S + ND}$$


Concept C1
Concept – Concept & Rel – Rel similarity
Concept C2
Sim(C1,C2) - % of shared ancestors
Predication – Predication similarity
Predication P1
Predication P2
Sim(P1,P2) - average pairwise similarity of subject, predicate, and object
Predication Set – Predication Set similarity
Predication Set Set1
Predication Set Set2
Sim(Set1, Set2) - average of maximum pairwise predication similarity
Document D1
Document – Document similarity
Document D2
Sim(D1, D2) – predication set similarity of the two documents

Document – Document similarity

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

0.5151

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

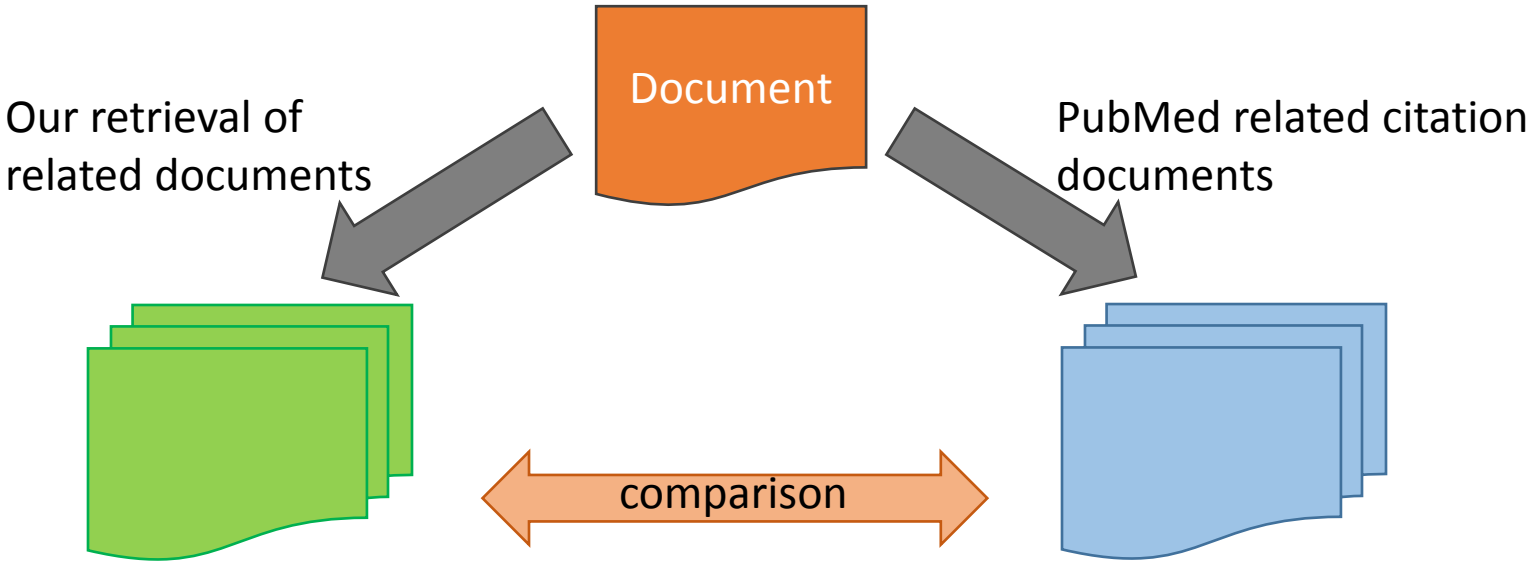
Predication set

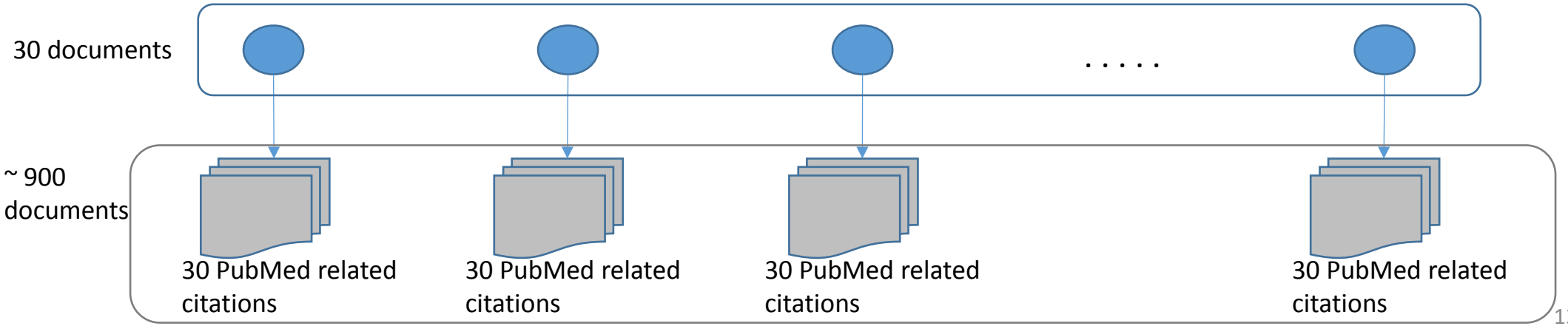
Entire intestinal epithelium	part of	Rattus norvegicus
Ileal mucous membrane	part of	Rattus norvegicus
Bacteria	location of	Lipopolysaccharides
Bacterial Translocation	process of	Animals
Endotoxins	causes	Systematic inflammatory Response Syndrome
……		
…		

Predication set

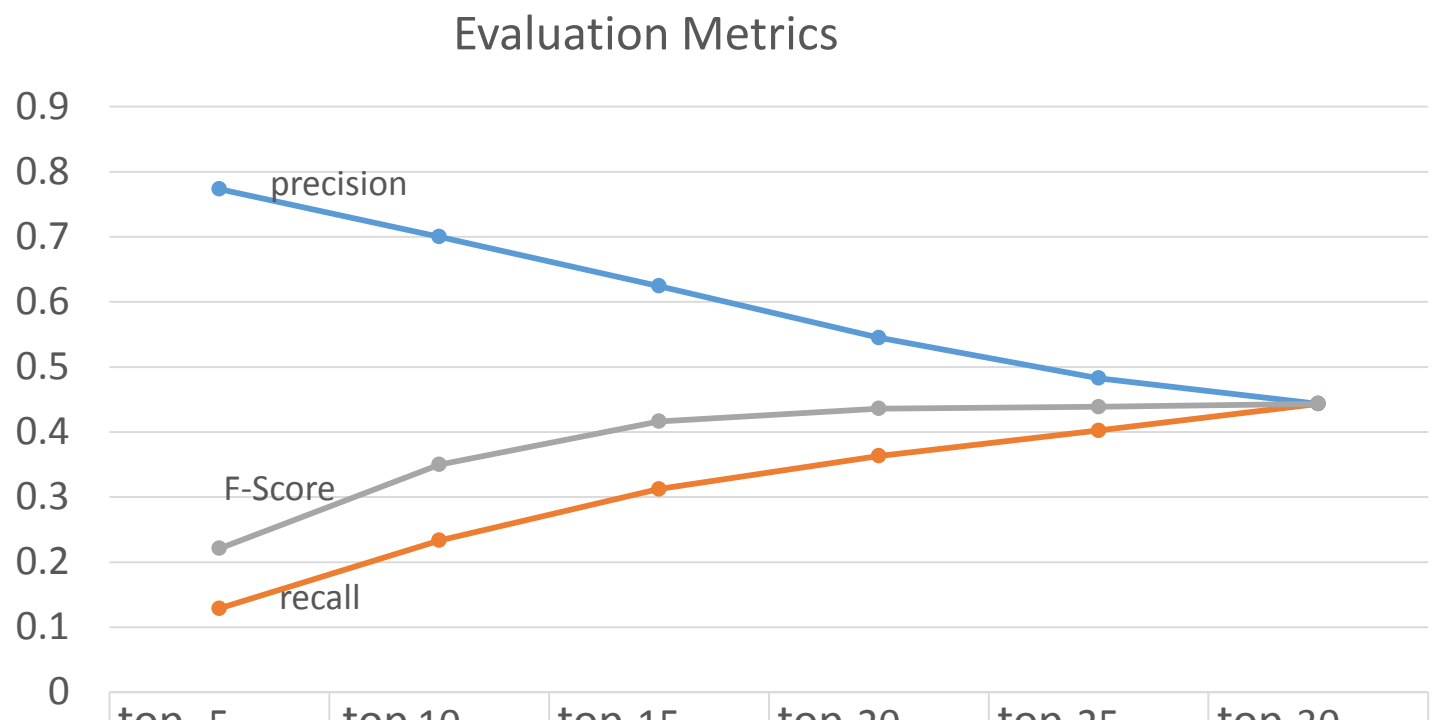
Large intestine	part of	Equus caballus (horse)
Nerve Fibers	part of	Submucous plexus
Tissue fiber	part of	Mucous membrane
Cell body neuron	location of	Nitric oxide
…..		
….		

Principle of our evaluation





Preliminary evaluation



	top 5	top 10	top 15	top 20	top 25	top 30
Precision	0.7733	0.7	0.6244	0.5449	0.4827	0.4433
Recall	0.1289	0.2333	0.3122	0.3633	0.4022	0.4433
F-Score	0.221	0.35	0.4163	0.436	0.4388	0.4433

retrieved documents

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 ?