

# 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

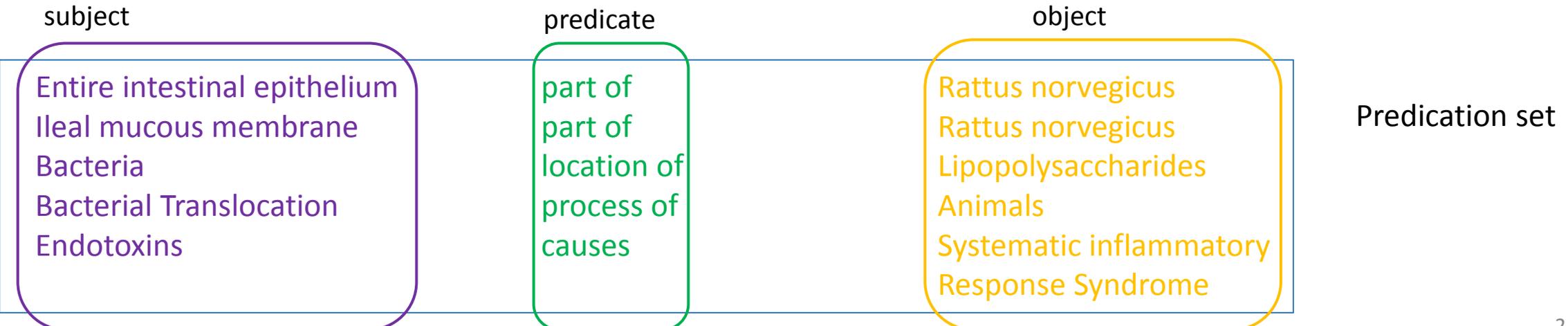
08.07.2014



# 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.



# 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

NCBI Resources How To Sign in to NCBI

PubMed.gov US National Library of Medicine National Institutes of Health

PubMed 9688090 Search

RSS Save search Advanced Help

Display Settings: Abstract

Shock. 1998 Jul;10(1):43-8.

**Pure endotoxin does not pass across the intestinal epithelium in vitro.**

Benoit R<sup>1</sup>, Rowe S, Watkins SC, Boyle P, Garrett M, Alber S, Wiener J, Rowe MI, Ford HR.

Author information

**Abstract**

Numerous reports suggest that endotoxin (LPS) may play a central role in triggering the inflammatory cascade that leads to the systemic inflammatory response syndrome. Although conditions that promote bacterial translocation in vivo may also facilitate direct translocation of LPS, the exact mechanisms by which LPS crosses the intestinal barrier to reach the systemic circulation are unknown. This study was designed to determine whether pure endotoxin could pass across injured rat ileal mucosa in the Ussing chamber. Sprague-Dawley rats were subjected to mild or severe hemorrhagic shock following carotid artery cannulation, and then resuscitated. Control animals underwent carotid artery cannulation only (sham-shock). Bacterial translocation to the mesenteric lymph nodes, liver, or spleen was measured after 24 h. Transmucosal passage of fluorescein isothiocyanate (FITC)-labeled E. coli C-25, or FITC-conjugated LPS was measured in the Ussing chamber. Intestinal membranes were examined by light and confocal laser microscopy. Severe hemorrhagic shock resulted in a 60% mortality rate and a 100% incidence of bacterial translocation in surviving animals. Sham-shock rats had a 100% survival rate and a 33% incidence of bacterial translocation. Transmucosal passage of FITC-E. coli C-25 was similar in both groups; however, passage of FITC-LPS was never detected. Histologic analysis confirmed mucosal injury to the intestinal epithelium of rats subjected to severe hemorrhagic shock, and confocal laser microscopy demonstrated passage of FITC-E. coli C-25, but not of

Send to:

Save items

Add to Favorites

**Related citations in PubMed**

Transmucosal passage of bacteria across rat intestinal epithelium in the Ussing chamber [Shock. 1998]

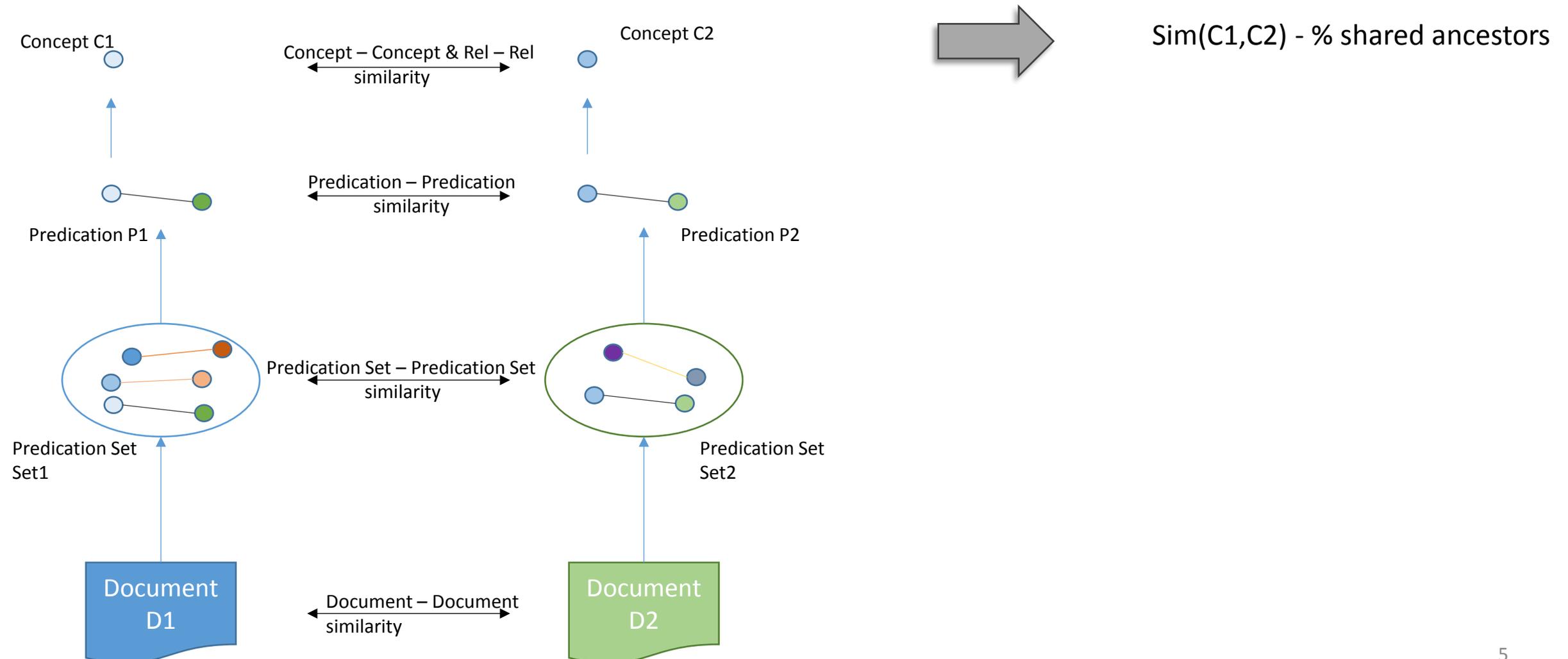
Inhibition of nitric oxide with aminoguanidine reduces bacterial translocation [Arch Surg. 1996]

The effect of endotoxin on intestinal mucosal permeability to bacteria in vitro. [Arch Surg. 1995]

Review [Induction mechanism of shock: applying the etiologic model] [Nihon Hoigaku Zasshi. 2004]

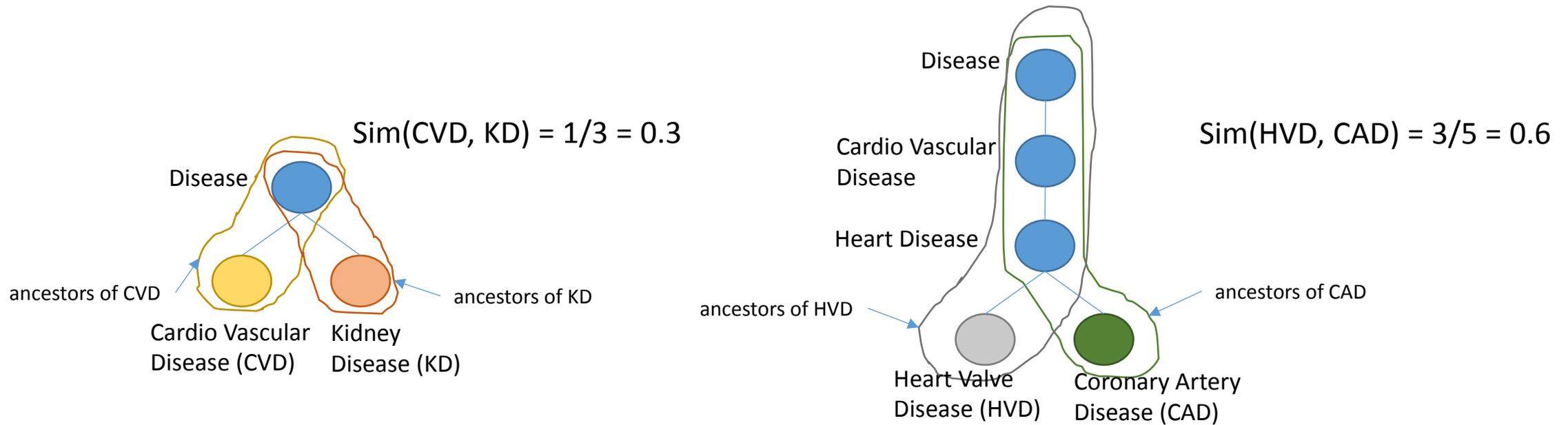
Review Clinical significance of translocation. [Gut. 1994]

See reviews...

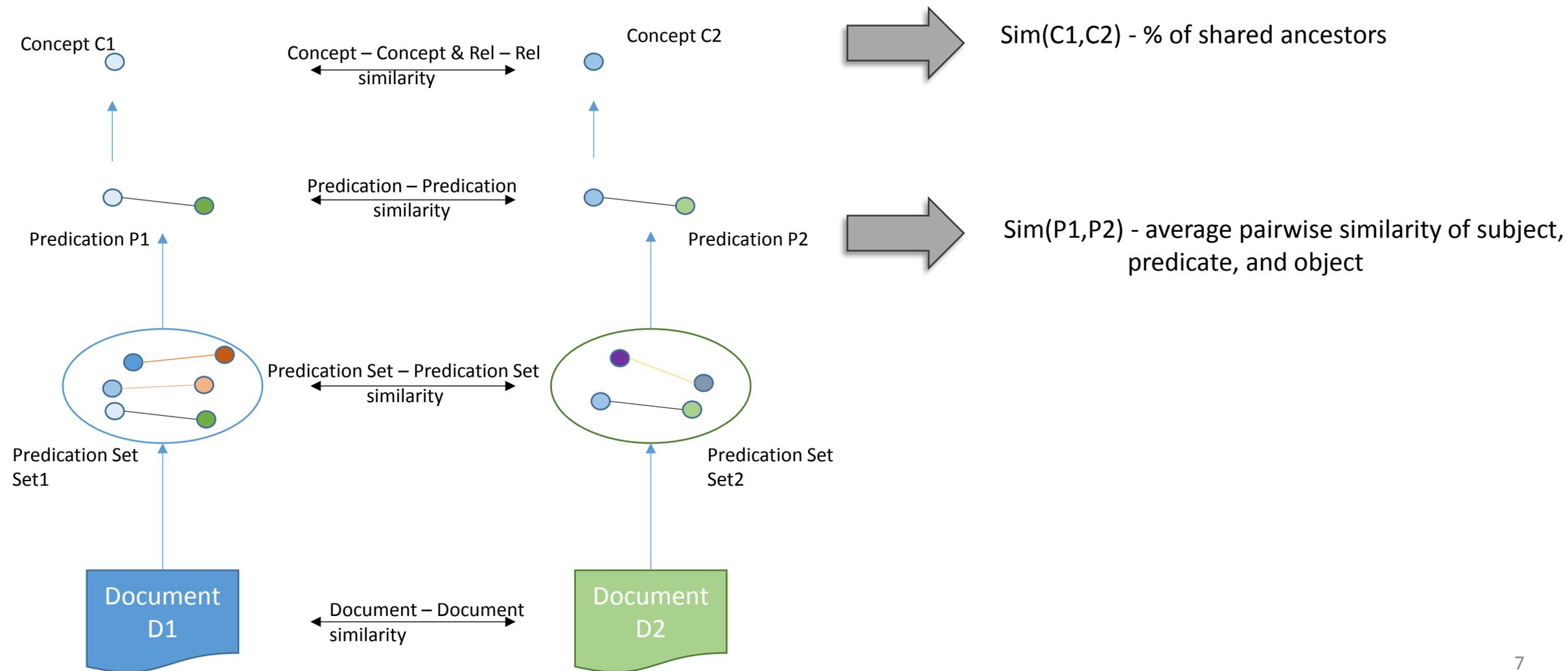


# Concept-Concept similarity

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

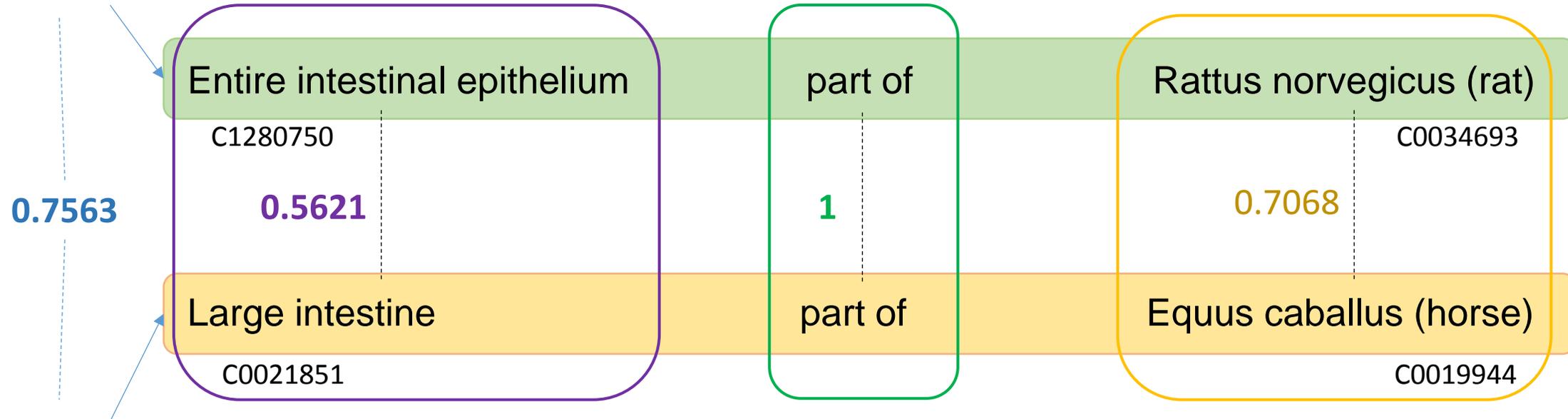


- Jaccard ( $c1, c2$ ) =  $\frac{\text{number of shared concepts between } c1 \text{ and } c2}{\text{total number of concepts in } c1 \text{ and } c2}$



# Predication-Predication similarity

Predication P1

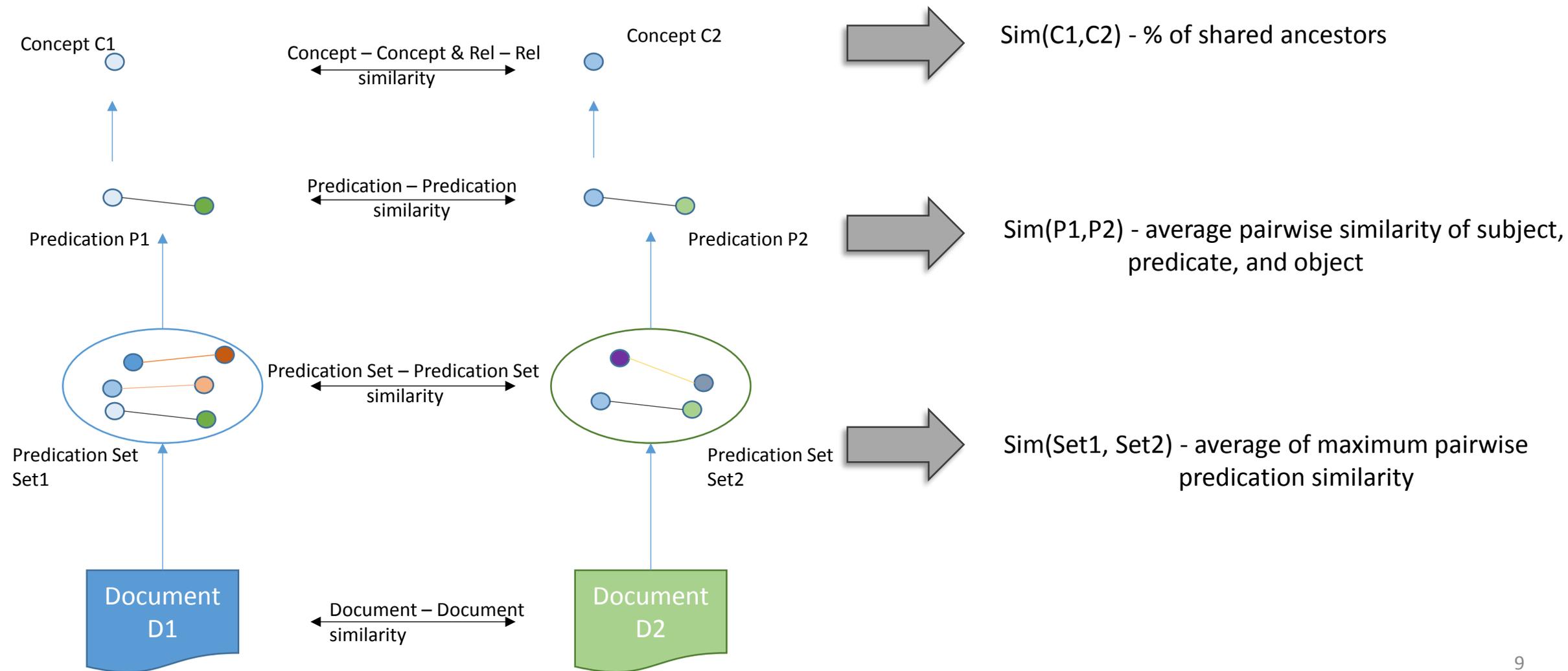


Predication P2

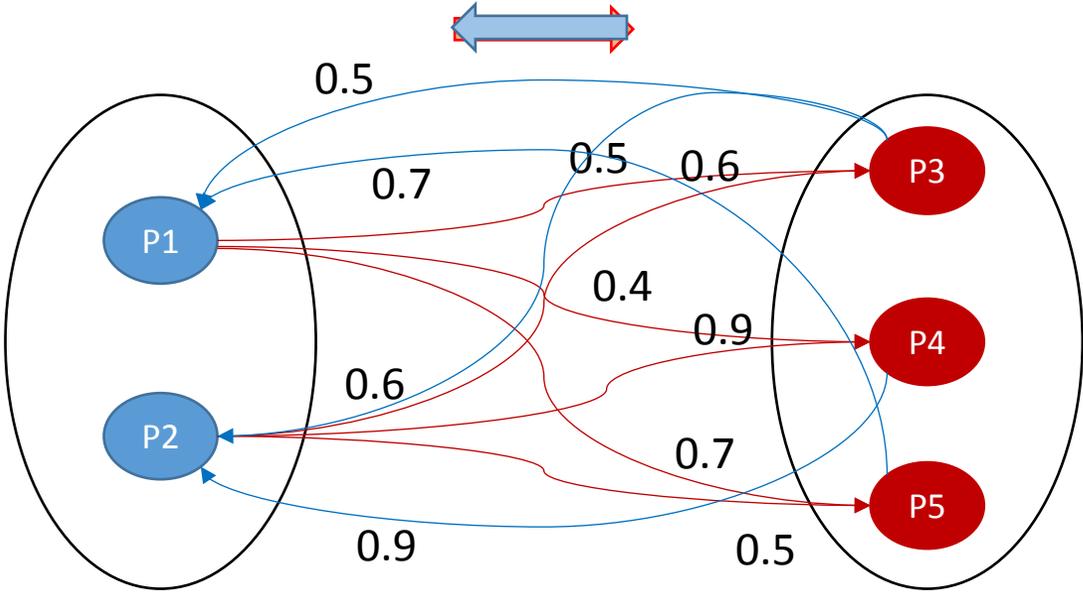
$$\text{Sim}(P1, P2) = W_s * \text{Sim}(C1, C2) + W_p * \text{Sim}(R1, R2) + W_o * \text{Sim}(O1, O2) / (W_s + W_p + W_o)$$

when,  $W_s = W_p = W_o = 1$

$$\begin{aligned} \text{Similarity} &= (0.5621 + 1 + 0.7068) / 3 \\ &= 0.7563 \end{aligned}$$



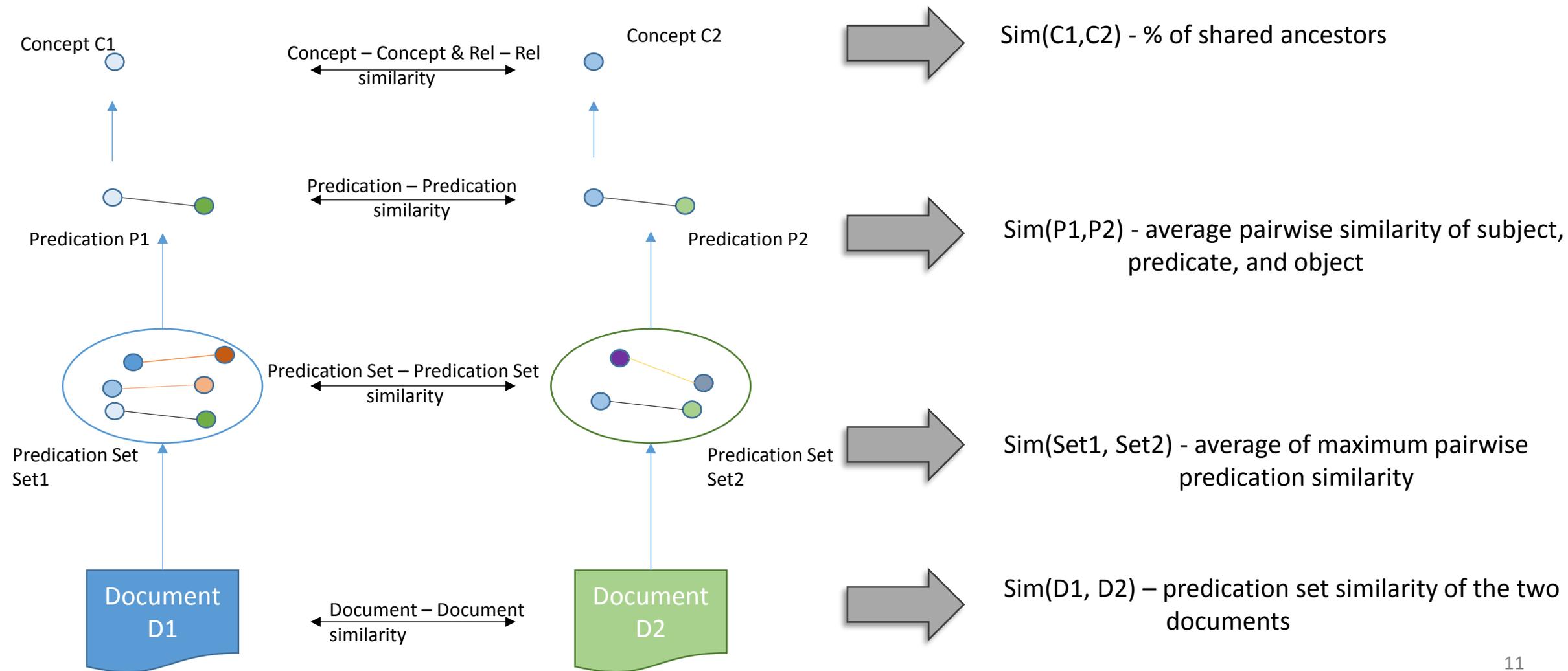
# Predication Set – Predication Set similarity



$$\text{Similarity} = (0.7 + 0.9 + 0.6 + 0.9 + 0.7) / (2 + 3)$$

$$= 0.76$$

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



# Document – Document similarity

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

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

0.5151

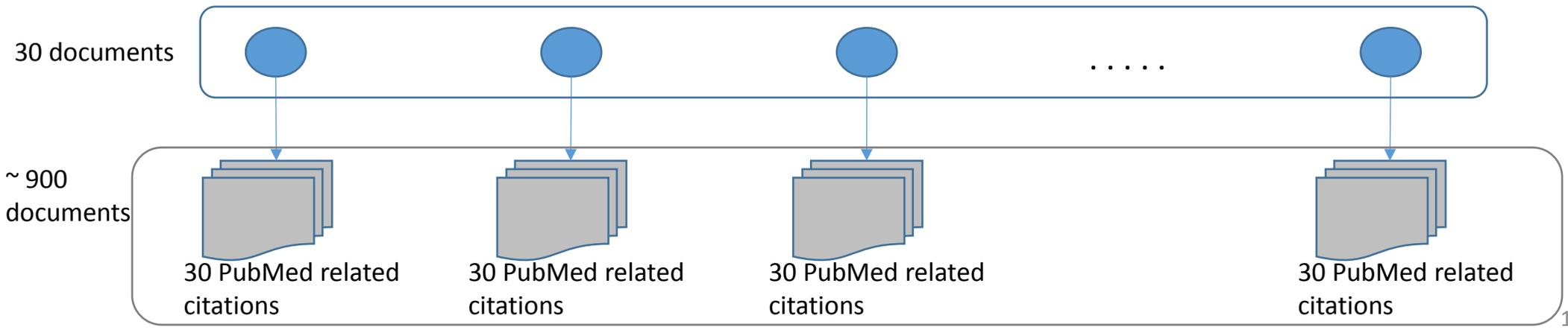
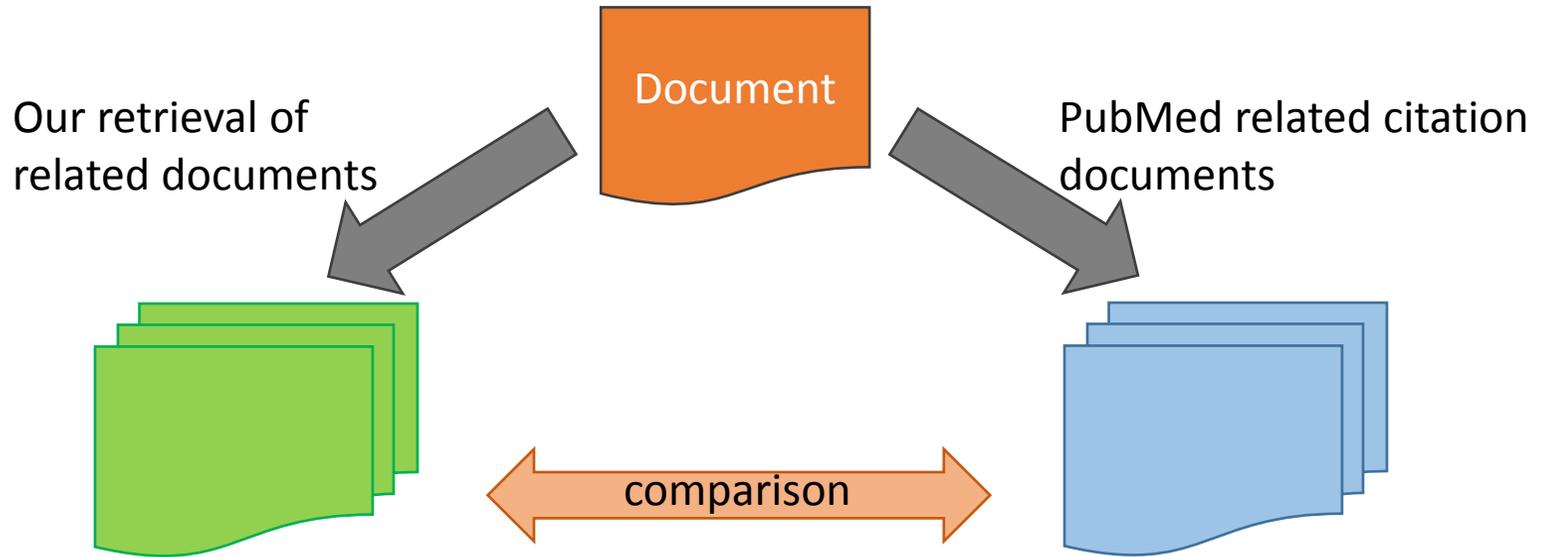
Predication set

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

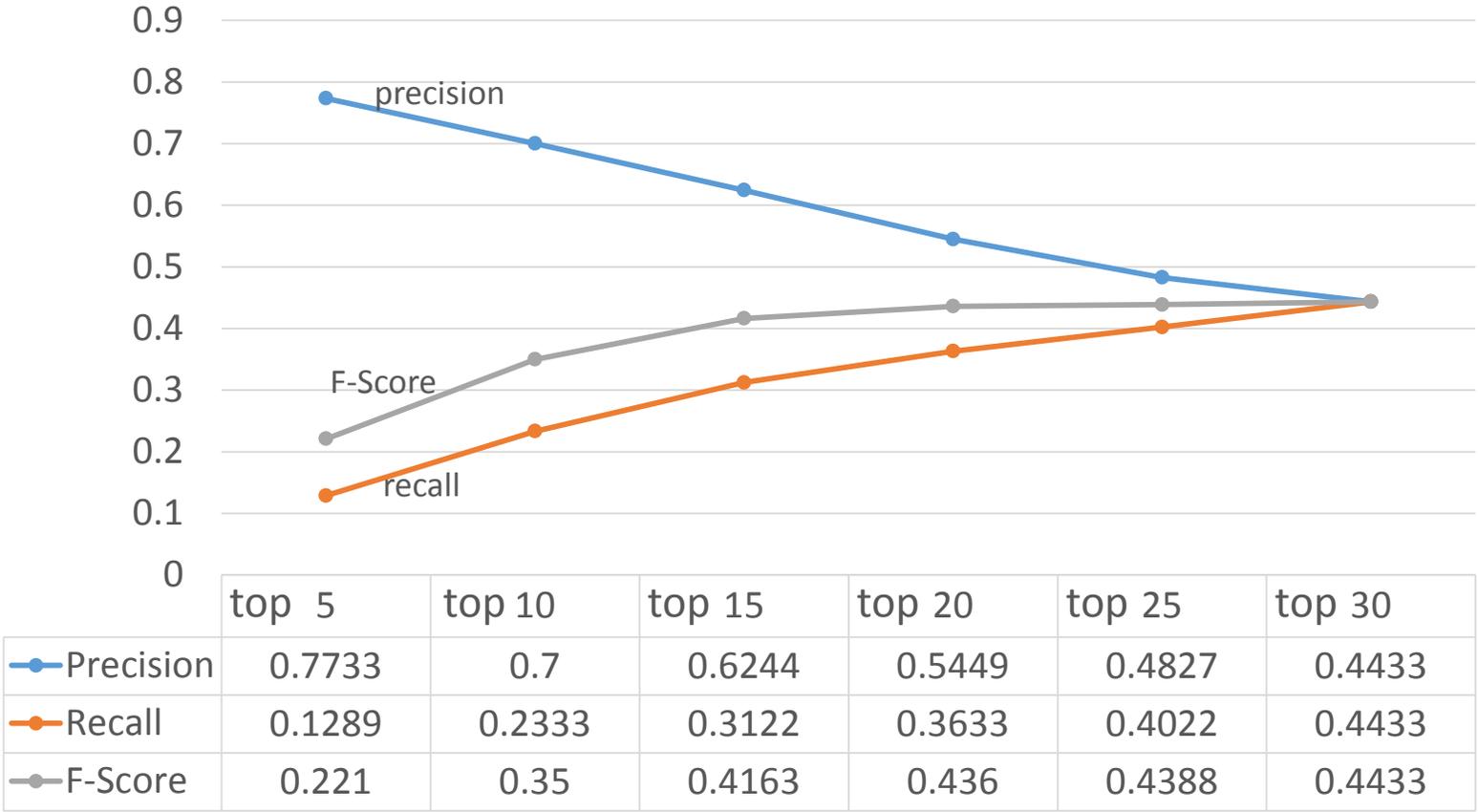
Large intestine	part of	Equus caballus (horse)
Nerve Fibers	part of	Submucous plexus
Tissue fiber	part of	Mucous membrane
Cell body neuron	location of	Nitric oxide
.....		
...		

# Principle of our evaluation



# Preliminary evaluation

Evaluation Metrics



# 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 ?