FAst Cheap and Synthetic Oracle (FACSO)

Proximity Measures to Capture Expert Knowledge in the "Bibliome"

Luis M. Rocha[†‡] and **Andreas Rechtsteiner**[†]

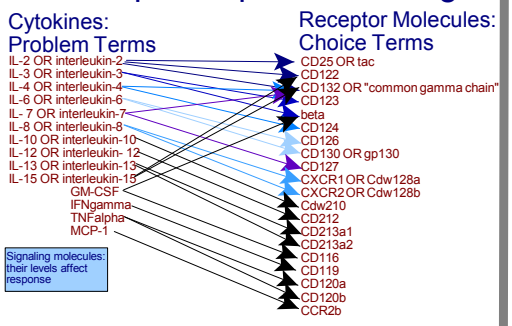
[†]CCS3 - Modeling, Algorithms, and Informatics, *Los Alamos National Laboratory*, MS B256,Los Alamos, NM 87545, USA

[‡]*Instituto Gulbenkian de Ciência*, 2781-901, Oeiras, Portugal

Abstract

The present-day existence of very large electronic collections of documents, particularly in Biomedicine, allows simple information retrieval (IR) statistical methods to automatically capture expert knowledge with very good accuracy. Here we presenta search utility (FACSO) built at Los Alamos, which queries the *Altavista*™ search engine to determine associations between two classes of biological objects in Immunology (though applicable to any domain). This search utility uses measures of proximity built from the co-occurrence of keywords in documents. We present the accuracy of this utility, and constrast it to the Pointwise Mutual Information - IR method. A new version of this tool using PubMed is now under development.

1. Example: Expert Knowledge



2. Co-Occurrence Proximity

Given a binary relation A between sets of keywords K and documents D we extract a co-occurrence proximity measure: $KDP(k_i, k_j)$ is the probability that both keywords k_i and k_j co-occur in the same document $d \in D$.

$$kdp(k_i,k_j) = \frac{\sum_{k=1}^{m}(a_{i,k} \wedge a_{j,k})}{\sum_{k=1}^{m}(a_{i,k} \vee a_{j,k})} = \frac{N_{\cap}(k_i,k_j)}{N_{\cup}(k_i,k_j)}$$

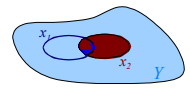
(Keyword Document Proximity)

$$KDP(k_i,k_j) = \frac{1}{\frac{1}{P_K(k_j|k_i)} + \frac{1}{P_K(k_i|k_j)} - 1}$$

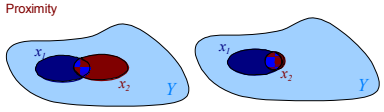
Where

$$P_K(k_i|k_j) = \frac{\sum_{k=1}^{m}(a_{i,k} \wedge a_{j,k})}{\sum_{k=1}^{m}(a_{j,k})} = \frac{N_{\cap}(k_i,k_j)}{N(k_j)}$$

Conditional Probability



Proximity



3. Altavista Queries

3.1 Proximity Queries

$$prox_1(problem, choice_i) = \frac{hits(problem \text{ AND } choice_i)}{hits(problem \text{ OR } choice_i)}$$

$$prox_2(problem, choice_i) = \frac{hits(problem \text{ NEAR } choice_i)}{hits(problem \text{ OR } choice_i)}$$

NEAR: co-occurrence within 10 words

$$prox_3(problem, choice_i) = \frac{hits((problem \text{ NEAR } choice_i) \text{ NEAR } context)}{hits(problem \text{ OR } choice_i)}$$

$context$ = {receptor}

3.2 Conditional Probability Queries

Turney, P.D. (2001). "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL." *Proceedings of the Twelfth European Conference on Machine Learning* (ECML-2001), Freiburg, Germany, pp. 491-502. http://extractor.iit.nrc.ca/reports/ECML2001.html

$$score_1(choice_i) = \frac{hits(problem \text{ AND } choice_i)}{hits(choice_i)}$$

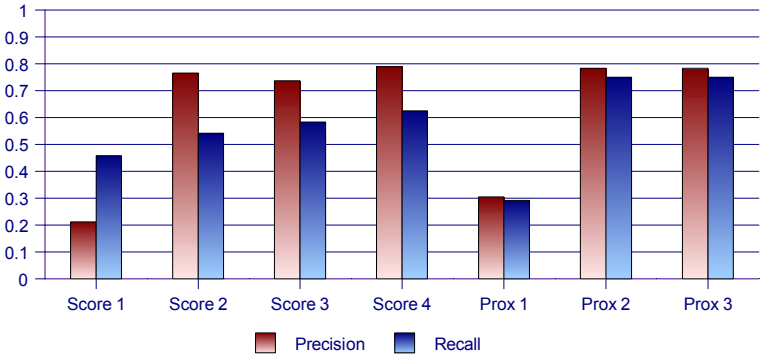
$$score_2(choice_i) = \frac{hits(problem \text{ NEAR } choice_i)}{hits(choice_i)}$$

$$score_3(choice_i) = \frac{hits(problem \text{ NEAR } choice_i) \text{ AND NOT}((problem \text{ OR } choice_i) \text{ NEAR "not"})}{hits(choice_i \text{ AND NOT}(choice_i \text{ NEAR "not"}))}$$

$$score_4(choice_i) = \frac{hits((problem \text{ NEAR } choice_i) \text{ NEAR } context \text{ AND NOT}((problem \text{ OR } choice_i) \text{ NEAR "not"}))}{hits(choice_i \text{ NEAR } context \text{ AND NOT}(choice_i \text{ NEAR "not"}))}$$

4. Relevant Associations

- Normalize proximity/score of choice terms for each problem term (Column Normalization)
- Calculate α-cut (α = 0.5)
  - Choice terms ≥ α are considered relevant associations



Precision: probability that an identified association is relevant

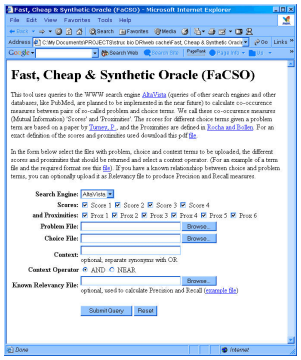
$$precision = \frac{\|\{relevant\} \cap \{retrieved\}\|}{\|\{retrieved\}\|}$$

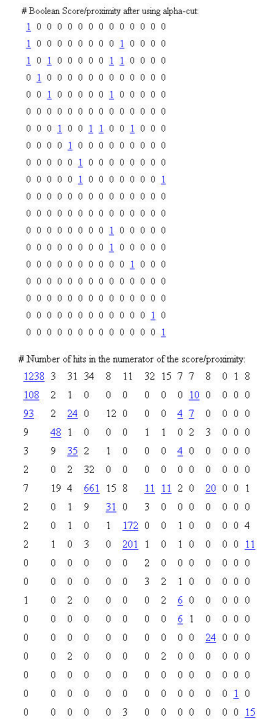
Recall: probability that an association has been identified given that it is relevant

$$recall = \frac{\|\{retrieved\} \cap \{relevant\}\|}{\|\{relevant\}\|}$$

{relevant} = {associations defined in 1}

5. Web Utility





- Discovers Relevant Associations
- Retrieves Documents substantiating the associations
- Being Developed for PubMed

http://bimmer.c3.lanl.gov/~andreas/easyfast8.html