Extraction and Semi-metric Analysis of Social and Biological Networks

Luis M. Rocha
CCS3 - Modeling, Algorithms, and Informatics, *Los Alamos National Laboratory*, MS B256,Los Alamos, NM 87545, USA

Abstract
We extract social networks from co-occurrence data incollections of electronic documents. These associative networks are represented as weighted graphs whose edges denote degrees of proximity or its inverse, a distance function. Most distance graphs obtained violate the triangle inequality expected of Euclidean distances. This type of distance function is known as a semi-metric. We show that the semi-metric behavior of these distance graphs, can be used for identifying specific implicit associations in the graph, and thus useful to identify trends in communities associated with the sets of documents from where associations were extracted.

1. Co-Occurrence Proximity

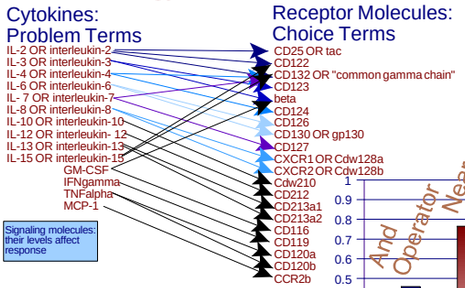
Given a binary relation A between sets of keywords K and documents D we extract a co-occurrence proximity measure: $KDP(k_i, k_j)$ is the probability that both keywords k_i and k_j co-occur in the same document $d \in D$.

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

(Keyword Document Proximity)

2. Power of Proximity Measures

Expert Knowledge: Immunology Testcase

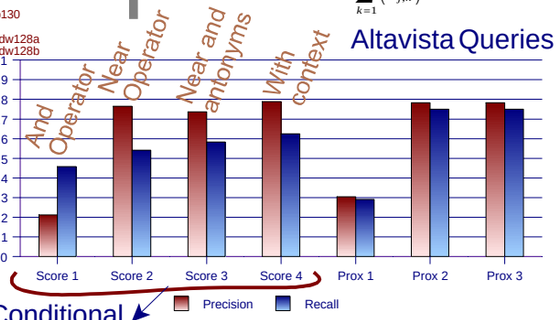


Where

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

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

Altavista Queries



Conditional Probability Queries

Precision: probability that an identified association is relevant

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

3. Distance Functions

$$d_{kdp}(k_i,k_j)=\frac{1}{kdp(k_i,k_j)}-1$$

(Keyword Document Distance)

d is a distance function because it is a nonnegative, symmetric, real-valued function such that $d(k, k) = 0$

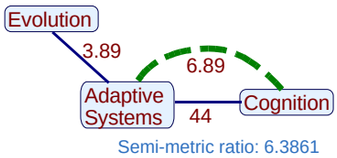
4. Semi-metric behavior

$$d(k_1, k_2) \le d(k_1, k_3) + d(k_3, k_2)$$

Metric

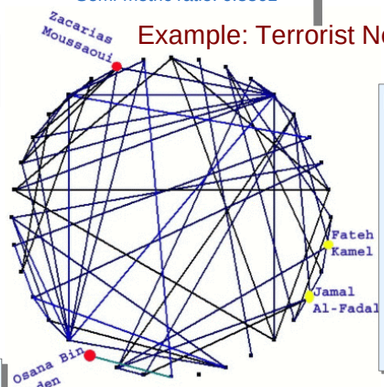
$$d(k_1, k_2) > d(k_1, k_3) + d(k_3, k_2)$$

Semi-metric



Semi-metric ratio: 6.3861

Example: Terrorist Networks



- Pairs with larger semi-metric behavior denote a *latent association*
  - Not grounded on direct evidence provided by the relation R, but rather implied by the overall network of associations in this relation.
  - Meaning depends on the semantics of the application
    - In graphs of keyword co-occurrence in documents: associated with novelty and can be used to identifytrends.
    - In terrorist networks identifies people, groups, etc. for which we do not have direct evidence, that a real association exists, but who could easily be indirectly associated.
  - In recommendation system for journals now at LANL
  - Applied to web sites, digital libraries, gene networks, random graphs

Semi-metric ratio

- Absolute measure of indirect distance reduction

$$s(x_i,x_j)=\frac{d_{direct}(x_i,x_j)}{d_{shortest}(x_i,x_j)}$$

Relative Semi-metric ratio

- Distance reduction against maximum contraction

$$rs(x_i,x_j)=\frac{d_{direct}(x_i,x_j)-d_{shortest}(x_i,x_j)}{d_{\max}-d_{\min}}$$

Below Average Ratio

- Captures semi-metric distance reductions which contract to below the average distance for a given node. Captures some of the cases of initial ∞ distance

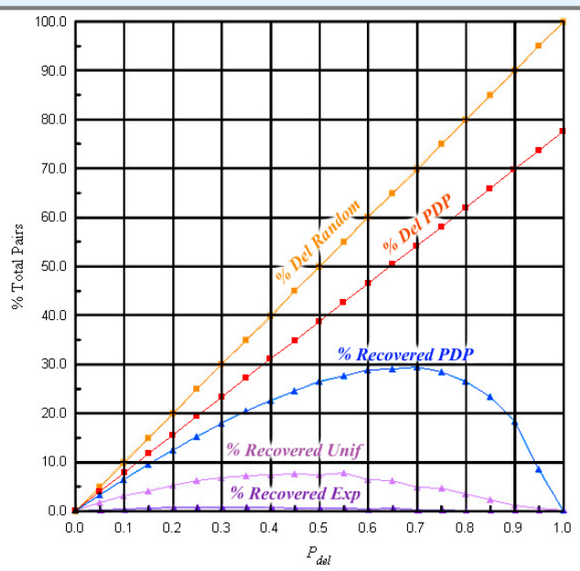
$$b(x_i,x_j)=\frac{\overline{d_{x_i}}}{d_{shortest}(x_i,x_j)}$$

Rocha, Luis M. [2002]. "Semi-metric Behavior in Document Networks and its Application to Recommendation Systems". In: *Soft Computing Agents: A New Perspective for Dynamic Information Systems*. V. Loia (Ed.) International Series Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 137-163.

5. Recovering Missing Knowledge

- Perfect Knowledge
  - Transitive Closure of real graph
  - Metric Distance Graph
- Incomplete Knowledge
  - Each positive association is deleted with probability p_{del}
  - 100 graphs for each value of p_{del}

Random Deletions (Full and Partial)



PDP against Random Graphs

Distance from People (terrorists) Document Proximity