

Graphical Models and Networks for Monolingual, Multilingual and Translingual Text Retrieval and Visualization

Mark Davis

madavis@crl.nmsu.edu

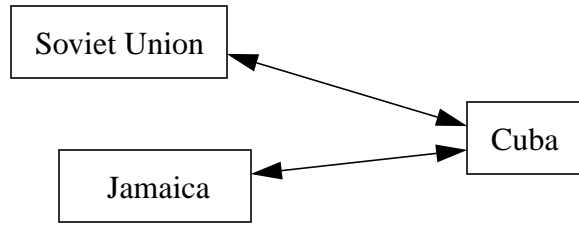
Computing Research Laboratory

New Mexico State University

Las Cruces, NM 88003

For a single text retrieval engine to support monolingual, multilingual and translingual text retrieval and visualization means that the engine must be built on a suitably general framework that uses simple feature co-occurrence for retrieval operations. The features must be more general than document-term occurrences, however, because a user may want to visualize more than the relationships between terms and documents. At Computing Research Laboratory, we have recently been examining a general class of graphical models for retrieval and visualization purposes. This marks a departure from our recent work on cross-language query optimization using corpus-based and lexical resources, but continues an avenue of research that was conducted here at CRL in the mid-1980s for explaining semantic relationships between terms. The methods are distinguished by their use of graph representations of term relationships and the reduction of the sparse graph data according to a simple accounting algorithm that erases node-node links in the graph when evidence weighs against a relationship between those nodes.

In the late 1960s, hierarchical, tree-like representations were proposed for explaining the way in which subjects classified objects (Quillian, 1969). The taxonomic system of biology is the obvious parallel to this sort of semantic arrangement; a dog is a type of mammal, which is a type of animal, and so forth. Problems quickly emerged in the use of hierarchies to represent semantic relationships, however, because the hierarchy imposed prototype-instance relationships that did not appear to match the relationships seen in actual subjects. For example, “canary” is closely associated with “yellow” in most subject’s minds, but not nearly so closely associated with “skin”, despite the fact that in a hierarchical model there is no favorable position granted to yellowness compared with skin from the position of the canary. By removing the requirement that the semantic network be hierarchical it became possible to represent frequency and uniqueness effects, and position objects more accurately in semantic space. Further, asymmetrical relationships can be captured by free network models, but not by trees. The classic example is that Cuba was closely related to the Soviet Union and Jamaica, but Jamaica is not closely related to the Soviet Union. This can be represented by bi-directional links in digraphs:



Asymmetrical relationships violate triangularity assumptions in semantic space and are therefore difficult to describe in geometrical interpretations of semantic spaces. A similar argument may be applicable to both vector-space models of information retrieval and Latent Semantic Indexing approaches, both of which assume that term-document relationships obey triangularity relationships and can therefore be handled in a geometric space by traditional distance metrics like Euclidean distance. These methods also make simple visualization of relationships difficult because the obvious Euclidean visualizations are of very high dimensionality and attaching meaning to the dimensions is often very difficult.

The primary graphical or network model that we are currently investigating is derived from the Pathfinder algorithm for explaining semantic proximity relationships. In a Pathfinder network, the r -metric is substituted for Euclidean distances (Schvaneveldt, Durso and Dearholt, 1987):

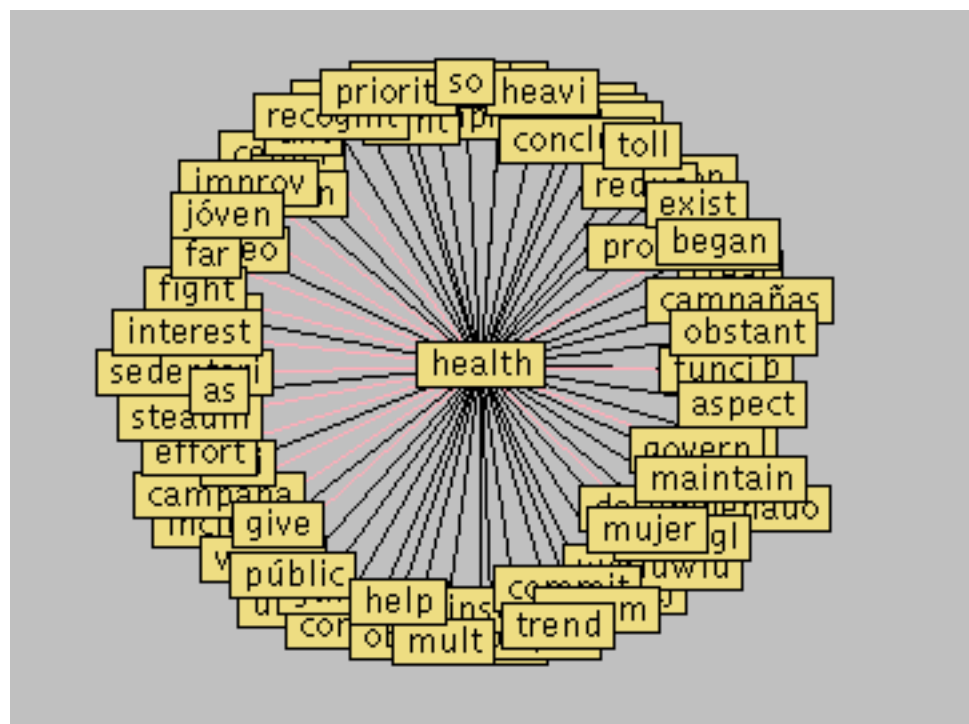
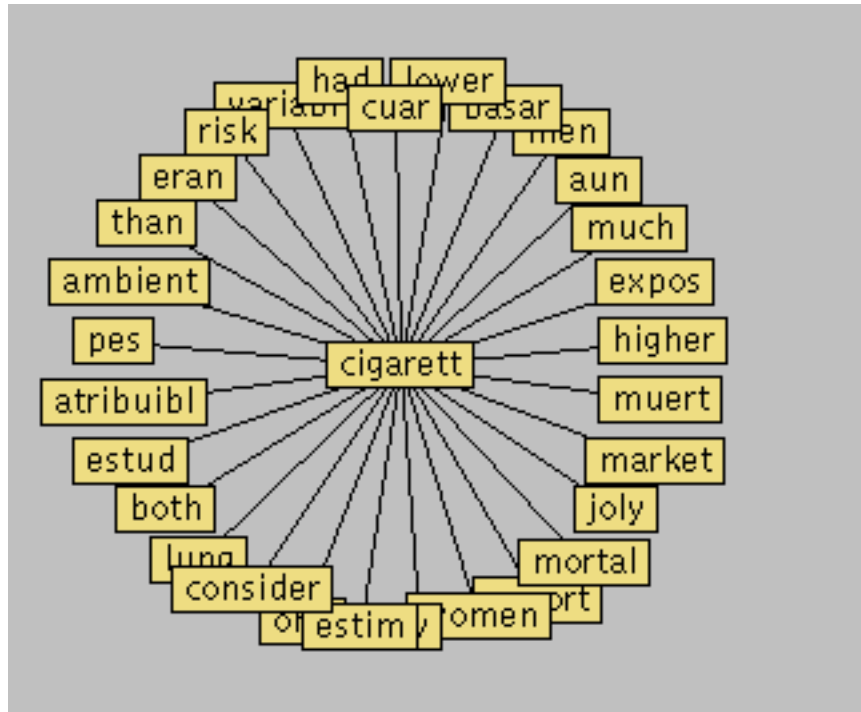
$$w(Path) = \left[\sum_k w_i^r \right]^{\frac{1}{r}} \quad (EQ 1)$$

where $w(Path)$ is the weight of a path in the graph consisting of k links with weights w , and $1 \leq r \leq \infty$. If $r = 1$, then the algorithm discards a direct node-node relationship if there exists a path between the nodes that has a greater or lesser length, derived from the simple sum of the weights along the path. The algorithm can also be parameterized in terms of how many links to traverse in a given path. Interested readers are directed to Schvaneveldt (1989) and Schvaneveldt, Durso and Dearholt (1987) for further discussion of the relationship between r and link length (“ q ”) in Pathfinder networks.

For our purposes, the Pathfinder algorithm serves as a dimensionality-reduction algorithm similar to Latent Semantic Indexing approaches. The proof is due to Ted Dunning (1997) and derives from considering the Pathfinder algorithm as an abstract form of matrix multiplication on text-cooccurrence data. The matrix resulting from the algorithm is then approximately determined by the dominating eigenvalues in the original data.

Monolingual retrieval using Pathfinder networks is accomplished by computing the network over a term by term plus document matrix. The term-term values can be calculated using any convenient cooccurrence measure like Average Mutual Information or log-likelihood statistics (we use the Dice coefficient). Term-document values can be computed using IDF values:

For translingual and multilingual text, the text collection can be supplemented with examples of parallel or comparable text. For English and Spanish retrieval of the opposite language from the query terms, we must have previously folded English and Spanish parallel texts into the network structure. Example networks are shown below illustrating the proximal network structure around the terms “cigarette” and “health” in a collection of 166 aligned sentences from a Pan American Health Organization document (more that the terms are shown in stemmed form here).



For visualization purposes, network models have clear advantages over vector-space and Latent Semantic Indexing approaches to monolingual, multilingual and translingual text retrieval applications because, as shown here, the relationships between terms, other terms and documents are explicit in the representation.

A further avenue under investigation at NMSU is using graphical models for interactive retrieval tasks. By examining the trees that span query terms in a textual database, inappropriate senses of terms can be quickly eliminated by the user. A further step will involve applying similar techniques to translingual and multilingual settings.

References

Quillian, M.R. (1969) The teachable language comprehender: A simulation program and theory of language. *Comm. of the ACM*, 12, 459-476.

Schvaneveldt, R.W., Durso, F.T and Dearholt, D.W. (1987) *Pathfinder: Networks from Proximity Data*, Technical Report MCCA-87-90, Computing Research Lab, New Mexico State University.

Schvaneveldt, R.W. (1989) *Proximities, Networks and Schemata*, Technical Report MCCA-89-149, Computing Research Lab, New Mexico State University.