

Hubs of Knowledge: using the functional link structure in Biozon to mine for biologically significant entities

Paul Shafer¹, Timothy Isganitis¹ and Golan Yona¹

Keywords: database search, ranking

1 Introduction.

Existing biological databases support a variety of queries such as keyword or definition search. However, they do not provide any measure of relevance for the instances reported, and result sets are usually sorted arbitrarily or by features irrelevant to the query (e.g. in alphabetical order). This is clearly not ideal as one might need to scan through hundreds or thousands of matches before encountering the instance that is the most relevant, the most studied, or the most interconnected. Furthermore, there are many instances in biological databases that are partially annotated or completely uncharacterized. Even if biologically relevant to the search term, these objects will be overlooked by traditional search methods. However, the relations between these objects and other, better annotated objects may help identify their functions. This may in turn imply that these objects are indeed relevant to the search term.

We describe a system that builds upon the complex infrastructure of the Biozon database and applies methods similar to those of Google to rank documents that match queries. We explore different prominence models and study the spectral properties of the corresponding data graphs. We evaluate the information content in non-principal eigenspaces and test various scoring functions for combining the contributions from multiple eigenspaces. We also test the effect of similarity data and other variations, which are unique to the biological knowledge domain, on the quality of the results. The result sets are assessed using a probabilistic approach that measures the significance of coherence between directly connected nodes in the data graph. This model allows us, for the first time, to compare different prominence models quantitatively and effectively and to observe unique trends. Our study resulted in a working ranking system of biological entities that was integrated into Biozon at biozon.org.

2 Methodology

We view important or interesting instances in the result sets as those that are linked to many other important entities. We study four main spectral methods for assigning prominence values to nodes: Eigenvector Centrality, Hubs and Authorities [1], PageRank [2] and Hybrid Katz's Status [3]. Given a query, the analysis starts by defining the graph (or subgraph) of relevant documents from the Biozon database and their **adjacency matrix A**. Each method differently characterizes the connectivity within the data graph to derive a **connectivity matrix B** from the adjacency matrix. The spectral properties of the connectivity matrix are then analyzed by computing its eigenvectors. Each eigenvector is considered to be a possible assignment of prominence values to documents, where node u is assigned a prominence value equal to the u^{th} component of the eigenvector. The highest scoring nodes in the principal vector(s) are returned as potential significant matches.

¹Dep. of Computer Science, Cornell University, Ithaca, NY, USA. E-mail: golan@cs.cornell.edu

We explore several variations of these methods that test the effect of different weighting functions, and inclusion of similarity data in the adjacency matrix. We also compare different search strategies (local methods that use just the local set of relations, or global that are based on pre-computed values for the complete graph).

3 Results

While each of the prominence models is designed to generate result sets that are sorted based on relevance, it is hard to quantitatively evaluate the quality of these results on a large scale. One of our contributions is the introduction of an evaluation methodology. To evaluate the quality and effectiveness of all these models and their variations thereof we propose an objective probabilistic measure, *UROC*, that accounts for both the structure of the Biozon graph and the textual information contained therein. This measure quantifies the thematic unity within instance subgraphs, directed at detecting what we call “hubs of knowledge”.

Specifically, for each instance v in the result set we examine its set of direct neighbors (denoted as the subgraph G_v) and count how many of them match the query term. We then compute the probability to observe such a local graph by chance, and its pvalue. The quality of the result set \mathbf{R} can be evaluated as the total pvalue

$$Q(\mathbf{R}) = - \sum_{v \in \mathbf{R}} \log(pvalue(G_v))$$

To account also for the ordering of objects within the result set. Our method is a variation over the popular ROC measure. However, unlike the typical setting for this measure (which requires labeled data) we have quantitative data with a significance value assigned to each sample. We assume that better model will report the more significant instances first. Therefore, the cumulative area under the curve corresponding to the sorted list of instances (from most significant to least significant) can serve as an overall performance measure, which we call UROC (unsupervised ROC).

We examine several issues with prominence models that have not been quantitatively addressed so far. We evaluate the utility of information contained in non-principal eigenspaces as well as different ways to incorporate this information into our prominence models. We observe that PageRank tends to consolidate the most significant information from across the other methods’ eigenspaces into its principal eigenspace. Practical considerations promoted the use of global methods for a real-time ranking system. Since PageRank’s principal eigenspace tends to incorporate information from the other methods’ non-principal eigenspaces, we conclude that the PageRank model with the principal eigenspace scoring function is the most feasible system for ranking inter-related query results. Our result suggest that sparse models (such as Eigenvector Centrality, Hubs and Authorities and Hybrid Katz’s Status) are most effective for producing hubs of knowledge (each one corresponding to a different eigenspace) while PageRank is the most effective and efficient model for ranking query results.

References

- [1] Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. *In 9th ACM-SIAM Symposium on Discrete Algorithms*
- [2] Page, L. Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web.
- [3] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika* **80**, 39-43.