Large-Scale Network Analysis to Improve Retrieval in the Biomedical Domain

Jimmy Lin
jimmylin@umd.edu
October 24, 2007

Note: This is a problem in search of a team leader and team members. I obviously can't run the course and be a team leader at the same time!

1. Problem

MEDLINE is the authoritative repository of abstracts from the primary literature in the biomedical domain, maintained by the National Library of Medicine. It currently contains over 17 million records. PubMed [1] is a publicly accessible Web gateway to the system. It is considered one of the most important resources available to researchers in biology, medicine, biochemistry, etc. Currently, PubMed implements a Boolean search algorithm that sorts results in reverse chronological order.

This proposed project will focus on developing better retrieval algorithms for MEDLINE by incorporating information from three sources:

- **Co-authorship network:** nodes in this network represent authors, and links between nodes represent authors who have co-authored one or more articles (links are bidirectional).
- **Citation network:** nodes in this network represent articles, and directional links between nodes represent explicit citations.
- **Document-similarity network:** nodes in this network represent articles, and directional links represent similarity in terms of abstract content (automatically computed).

As conceived, the project has two explicit goals: 1.) to characterize the properties of these large-scale networks, and 2.) to leverage features of these networks to improve retrieval in the biomedical domain.

2. Resources

Through my collaboration with colleagues at the National Library of Medicine, I am very likely able to obtain access to a significant subset of MEDLINE (few million records), along with raw data for the networks discussed above.

3. The MapReduce Perspective

To my knowledge, no one has analyzed the types of networks discussed above in the biomedical domain. As a first step, one might apply PageRank to the three networks. This would, for example, tell us who the central authors in the collection are and which papers are seminal. This information could then be integrated into a retrieval algorithm.

Another possibility includes applying algorithms for finding community structure in networks [2]. Communities might represent, for example, coherent sub-disciplines, researchers using the same methodology, etc. This information could be exploited to disambiguate queries, for example.

One interesting question is the extent to which these networks overlap. I anticipate that areas where they *don't* overlap will be a rich source of interesting discoveries. For example, it's probably the case that authors cite other articles similar in content, but it's worth looking at the cited articles that are very different in content. As one possibility, they might be articles that bridge different fields. The ability to identify such work would be important for information retrieval, literature discovery, etc.

4. Interesting Extensions

In addition to MEDLINE, the National Library of Medicine also maintains other collections of information: databases ranging from semi-structured gene records to actual sequences. There exist links that span databases: for example, gene records link to MEDLINE articles that discuss the particular gene.

An interesting extension of this work would be to expand network analysis to inter-database connections. Imagine being able to search for an article about a gene, and have the system automatically draw connections to other interesting articles (via multiple databases): for example, finding the database record of the gene, finding homologs of that gene in different organisms, and retrieving articles pertaining to the related gene.

5. Acknowledgements

Many of the ideas in this document originated from Michael DiCuccio at the National Library of Medicine.

6. REFERENCES

[1] http://www.ncbi.nlm.nih.gov/sites/entrez

[2] Girvan, Michelle and Mark E. J. Newman. (2002) Community Structure in Social and Biological Networks. Proceedings of the National Academy of Science, 99(12):7821-7826.