GenIRL
Genomic Information Retrieval using links

Stephen Blott Fabrice Camous Cathal Gurrin
Gareth J. F. Jones Alan F. Smeaton

School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland

{sblott, fcamous, cgurrin, gjones, asmeaton}
@computing.dcu.ie

1. INTRODUCTION

Given the diversity of biological information, we strongly believe that a retrieval system can perform better by integrating the links that exist between biological databases covering different areas and different types of data. As biologists identify new genes and gene functions every day, new sequences are stored and new literature is published at an increasing speed. The size of nucleotide sequences databases such as GenBank is growing larger as well as the size of protein sequences, protein structures and biomedical articles databases. The data is often structured and organized according to a project covering a specific area, e.g. a specific model organism. It is therefore difficult for biologists to find information that is not directly related to their field of research. Fortunately, much work was done on linking the various databases by annotating the records with references to external database records. These records and links form a complex graph that takes a lot of time and effort for users to navigate and search. Anybody who has searched the web is familiar with the frustrating experience of pursuing links, backtracking, and returning to the search engine to reformulate the query and begin again. We clearly need to find better ways to search and navigate the biological information in an integrated fashion.

2. GenIRL

GenIRL (Genomic Information Retrieval using Links) is a research project at Dublin City University that is investigating the usefulness of links in the information retrieval of genomic data. Various forms of links are being explored. Links can be extracted from textual similarity between publications or the presence of citations in scientific articles. They can underline sequence similarity or the homologous relation between genes. These links can also go across different domains, as in the case of a sequence being associated with a publication. Figure 1 illustrates some of the link types that we just mentioned. The goal of our project is to build a graph of this biological hypermedia which contains many heterogeneous link types. This graph will then be integrated in our retrieval system similarly to the way Google uses the PageRank algorithm to retrieve highly relevant web pages [2][3][7][8].

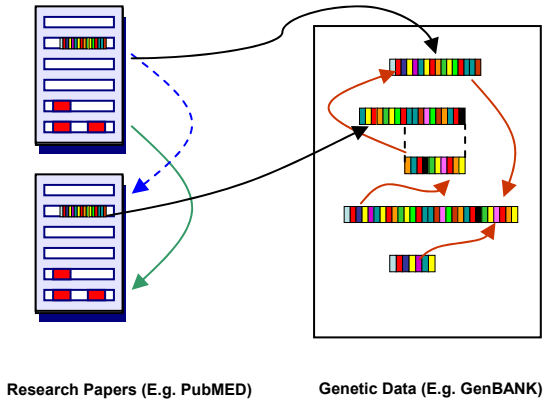

Figure 1. An example of the various types of links: citations between publications (green arrow), sequence similarity within publication databases (blue dashed arrows), sequence similarity within sequence databases (red arrows) and sequence similarity across different databases (black arrows).

We plan to use the genomic track of TREC (Text REtrieval Conference) as a vehicle to drive the experimental aspect of our project. The TREC guidelines and common evaluation procedures allow research groups from all over the world to evaluate their progress in developing and enhancing information retrieval systems. TREC supports experiments into different aspects of information retrieval with different tracks introduced since the establishment of TREC in 1992. Last year was the first Genomic track year and we participated in the ad hoc retrieval task during the Summer of 2003. The task required us to run 50 queries against more than 550,000 documents from Medline, the US National Library of Medicine (NLM) database of indexed journal citations and abstracts. A Medline record includes title, abstract and annotation fields such as Medical Subject Headings or MeSH terms, a controlled hierarchical vocabulary. A simplified view of the MeSH hierarchy is shown in figure 2.

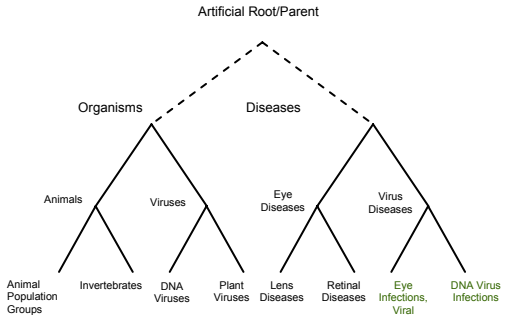

Figure 2. A simplified representation of the MeSH hierarchy. This hierarchy actually consist of 15 trees and there can be more than two children nodes per parent. An artificial root for all the trees was added for our

Our approach was motivated by the discovery and integration of links in the retrieval process. Most documents in the collection are annotated with several MeSH terms. We used the hierarchy to compute a measure of the similarity between each term. A similarity measure between each document could then be calculated by combining the similarity measures between the MeSH terms each document contained. The collection can be visualized as a graph where the nodes are the documents and the links represent a measure of the MeSH term-term similarity between the documents. The links of the graph are used in the retrieval and ranking of documents. We experimented with the use of this term-term similarity as a way to improve retrieval [1] and we chose to work on results generated using the SMART-based Okapi approach provided by Jacques Savoy of the University of Neuchatel, Switzerland. We used pseudo-relevant feedback, extracting MeSH terms from alleged relevant documents and building a new query according to Robertson's Offer Weight method [9]. The result of our official run showed an increase of 2.08 % in Mean Average Precision from the Okapi baseline (from 0.1635 to 0.1669).

Although we only obtained small improvements, the experiments were of an exploratory nature and current work is concentrating in several areas including investigation of alternative methods to compute MeSH term-term similarity measures [4].

3. WORK IN PROGRESS

Building on these first experiments, we aim at integrating more sources of links and given that linkage alternatives exist besides MeSH terms, we are exploring ways of using the Gene Ontology (GO) [5][6] in the forthcoming Genomic track of TREC 2004. The GO structure represents an immense source of links between gene products within species and also across species, and our goal is to integrate this structure in the search process in order to improve the performance of our Genomic retrieval system.

The structure of the GO vocabulary consists of three directed acyclic graphs. These represent information sets that are common to all living forms and are basic to the annotation of genes and gene products: 1) molecular function, 2) biological process and 3) cellular component. We are investigating different types of ways to build a graph with the document collection: documents can directly share the same GO term. They can also share related (parents, children within the GO hierarchy) GO terms. Besides, each GO term can be associated with a set of genes and gene products. The similarity of GO terms annotating two documents can then be measured according to the similarity of the genes and gene products they are associated with. The challenge here is to compute good measures for the links in order to build a graph that is a good model of the link structure of the document collection.

ACKNOWLEDGMENTS

GenIRL is funded by Enterprise Ireland under the Basic Research Grants Scheme, project number SC-2003-0047-Y.

REFERENCES

[1] Blott, S., Gurrin, C., Jones, G. J. F., Smeaton, A. F., Sodring, T. "On the use of MeSH headings to improve retrieval effectiveness". In Proceedings of the 12th TREC Conference, Gaithersburg, Md, November 2003.

[2] Brin, S., Page, L., "The anatomy of a large-scale hypertextual web search engine". Proceedings of the 7th WWW Conference, 1998.

[3] Chakrabarti, S.: 2003, "Mining the Web: Discovering knowledge from hypertext data". San Francisco: Morgan Kaufmann.

[4] Ganesan, P., H. Garcia-Molina, and J. Windom: 2003, "Exploiting Hierarchical Domain Structure to Compute Similarity". ACM Transactions on Information Systems 21(1), 64-93.

[5] Gene Ontology Consortium, "Creating the Gene Ontology Resource: Design and Implementation". Genome Research 11:1425-1433 (2001).

[6] Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology". Nature Genetics 25:25-29 (2000).

[7] Menczer, F.: "Combining Link and Content Analysis to Estimate Semantic Similarity". Proc. 13th Intl. WWW Conf. Alt. Track Papers and Posters, pp. 452-453, 2004

[8] Menczer, F.: "Lexical and Semantic Clustering by Web Links". To appear in JASIST.

[9] Robertson, S. E. & Spark Jones, K. "Simple, proven approaches to text retrieval". Technical report 356, Computer Laboratory, University of Cambridge.