International Journal of Advanced Computer and Mathematical Sciences
ISSN 2230-9624. Vol 3, Issue 4, 2012, pp 413-417
http://bipublication.com

HIERARCHIES BASED QUERY RESULTS FROM PUBLIC MEDLINE DATABASE

Shasikala Channamallu, N. Sujata Kumari, K. C. Ravi Kumar

Dept of CSE, Sri Devi Women's Engineering College, Hyderabad, A.P, India

[Received-11/09/2012, Accepted-03/10/2012]

ABSTRACT:

When you perform searching on a biomedical database such, it often returns a large number of results. Comprising of only a small subset is relevant to the user. Categorization results for the biomedical databases are the focus of this work. Organizing biomedical citations according to their MeSH annotations is the natural way, which is used as comprehensive hierarchy by PubMed. A search engine that enables the users to navigate large number of query results by organizing them using the MeSH hierarchy, a concept called BioNav system. Initially the query results are organized into a navigation tree. BioNav enables users to navigate large number of query results by categorizing them using MeSH; a comprehensive concept hierarchy used by PubMed.

Indexing terms: PubMed, BioNav system, MeSH hierarchy

INTRODUCTION:

In this current paper, we are stating the article search and citations at sentence lelvel and web search for PubMed. PubMed query and related article search are among the most useful and effective information retrieval tools for biologists and health care professionals. It has been demonstrated that the related article search is an often-exploited feature of PubMed [31]. The relatedness between two articles is computed based on the word similarity of their articles, titles and MeSH annotations [32]. This method is very effective in finding related articles, as demonstrated by Lin et al. [31, 33]. On the average, an article *A* can have hundreds of related article links. The related links of *A* are usually ranked in order of their similarity scores.

Besides the PubMed method, co-citation analysis is another established approach for linking related information in literature [34, 35]. Existing citation based approaches mostly rely on co-citation frequency to single out related articles. For example, two articles *A* and *B* are considered to be related if they are frequently co-cited by other articles [36], or if they cite the same large set of other articles[37].

By default, PubMed returns 20 search results in a page and displays the title, abstract and other bibliographic information when a result is clicked. Recent studies focus on two kinds of extensions to the standard PubMed output. First, because a PubMed search typically results in a long list of citations for manual inspection, systems mentioned in ‘Clustering results into topics’ section aim to provide an aid with a short list of major topics summarized from the retrieved articles [39].

The MEDLINE database, on which the PubMed search engine operates, contains over 18 million citations, and the database is currently growing at the rate of 500,000 new citations each year [19]. Other biological sources, such as Entrees Gene [17] and OMIM [20], witness similar growth. As claimed in previous work [25], the ability to rapidly survey this literature constitutes a necessary step towards both the design and the interpretation of any large scale experiment. Biologists, chemists, medical and health scientists are used for searching their domain literature such as PubMed using a keyword search interface. Currently, in an exploratory scenario where the user tries to find citations relevant to their line of research and hence not known a priori, submits an initially broad keyword- based query that typically returns a large number of results. Subsequently, the user iteratively refines the query, if she has an idea of how to, by adding more keywords, and re-submits it, until a relatively small number of results are returned. This refinement process is problematic because after a number of iterations the user is not aware if he has over-specified the query, in which case relevant citations might be excluded from the final query result.

Disadvantages:

- The ability to rapidly survey this literature constitutes a necessary step toward both the design and the interpretation of any large scale experiment.
- This refinement process is problematic because after a number of iterations the user is not aware if she has over-specified the query, in which case relevant citations might be excluded from the final query result.

Proposed System:

We retrieved citing sentences and citation lists by processing the full-text articles provided by the PMC Open Access database. This database contains a set of about 100,000 full-text articles [38]. A query on PubMed for “cancer” returns more than 2 million citations. Even a more specific query for “prothymosin”, a nucleoprotein gaining attention for its putative role in cancer development, returns 313 citations. The size of the query result makes it difficult for the user to find the citations that she is most interested in, and a large amount of effort is expended searching for these results. Many solutions have been proposed to address this problem commonly referred to as information overload [1, 2, 3, 9]. These approaches can be broadly classified into two classes: ranking and categorization, which can also be combined.

BioNav belongs primarily to the categorization class, which is ideal for this domain given the rich concept hierarchies

(e.g., MeSH [18]) available for biomedical data. We augment our categorization techniques with simple ranking techniques. BioNav organizes the query results into a dynamic hierarchy, the *navigation tree*. Each concept (node) of the hierarchy has a descriptive label. The user then navigates this tree structure, in a top-down fashion, exploring the concepts of interest while ignoring the rest.

An intuitive way to categorize the results of a query on Pub Med is using the Mesh static concept hierarchy [18], thus utilizing the initiative of the US National Library of Medicine (NLM) to build and maintain such a comprehensive structure. Each citation in MEDLINE is associated with several Mesh concepts in two ways: (i) by being explicitly annotated with them, (ii) By mentioning those in their text. Since these associations are provided by Pub Med, a relatively straightforward interface to navigate the query result would first attach the citations to the corresponding Mesh concept nodes and then let the user navigate the navigation tree.

ADVANTAGES:

Many solutions have been proposed to address this problem commonly referred to as information overload. These approaches can be broadly classified into two classes: ranking and categorization, which can also be combined.

Modules Description:

- Navigation Model
- TOPDOWN Cost Model

Navigation Model:

After the user issues a keyword query, BioNav initiates navigation by constructing the initial active tree (which has a single component tree rooted at the Mesh root) and displaying its root to the user. Subsequently, the user navigates the tree by performing one of the following actions on a given component sub tree I(n) rooted at concept node n:

1. **EXPAND I**(n): The user clicks on the">>>" hyperlink next to node n and causes an Edge Cut(I(n))operation to be performed on it, thus revealing a new set of concept nodes from the set I(n).

2. **SHOWRESULTS I**(n): By performing this action, the user sees the results list L(I(n)) of citations attached to the component subtree I(n).

3. **IGNORE I**(n): The user examines the label of concept node n, ignores it as unimportant and moves on to the next revealed concept.

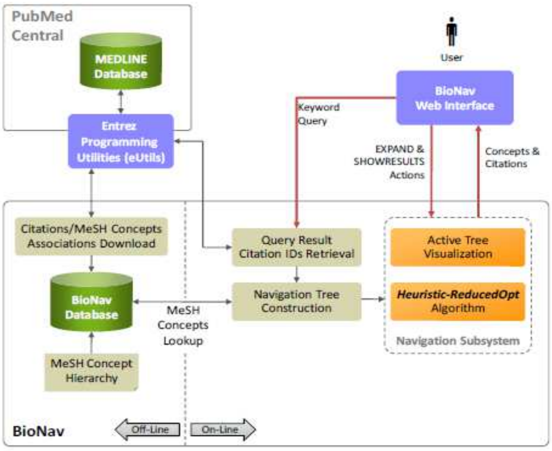
4. **BACKTRACK**: The user decides to undo the last Edge Cut operation.[13]

TOPDOWN Cost Model:

The cost model, which is inspired by a previous work [2], takes into consideration the number of concept nodes revealed by an EXPAND action, the number of EXPAND actions that the user performs and the number of citations displayed for a SHOW RESULTS action. In particular, the cost model assigns

- Cost of 1 to each newly revealed concept node that the user examines after an EXPAND action.
- Cost of 1 to each EXPAND action the user executes.
- Cost of 1 to each citation displayed after a SHOWRESULTS action.

System Architecture:[13]



Algorithm used:

- Optimal Algorithm for Best Edge Cut
- Heuristic-Reduced Opt Algorithm

CONCLUSION

In this current paper, we stated the article search and citations at sentence lelvel and web search for PubMed. In this paper, we presented the BioNav system to address this problem. Our solution is to organize the query results according to their associations to concepts of the MeSH concept hierarchy, and provide a dynamic navigation method that minimizes the information overload observed by the user. When the user expands a MeSH concept on our web interface, BioNav reveals only a selective list of descendant concepts, instead of simply showing all its children, ranked based on their estimated relevance to the user's query.

REFERENCES

[1] J S. Agrawal, S. Chaudhuri, G. Das and A. Gionis: Automated Ranking of Database Query Results. In Proceedings of First Bien-nial Conference on Innovative Data Systems Research (CIDR), 2003.

[2] K. Chakrabarti, S. Chaudhuri and S.W. Hwang: Automatic Categorization of Query Results. SIGMOD Conference 2004: 755-766.

[3] Z. Chen and T. Li: *Addressing Diverse User Preferences in SQLQuery-Result Navigation*. SIGMOD Conference 2007: 641-652

[4] L. Comtet: *Advanced Combinatorics: The Art of Finite and InfiniteExpansions, rev. enl. ed.* Dordrecht, Netherlands: Reidel, pp. 176-177, 1974.

[5] R. Delfs, A. Doms, A. Kozlenkov and M. Schroeder: *GoPubMed:Ontology-Based Literature Search Applied to Gene Ontology and PubMed*. German Conference on Bioinformatics 2004: 169-178.

[6] D. Demner-Fushman and Jimmy Lin: *Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering*. International Conference on Computational Linguistics and the Annual Meeting of the Association For Computational Linguistics, 2006: 841-848

[7] (2008) Entrez Programming Utilities. [Online].Available: http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

[8] U. Feige, D. Peleg and G. Kortsarz: *The Dense k-Subgraph Problem*. Algorithmica 29 (2001) 410-421

[9] V. Hristidis and Y. Papakonstantinou: *DISCOVER: Keyword Search in Relational Databases*. In Proc. of VLDB Conference, 2002

[10] R. Hoffman and A. Valencia: *A gene network for navigating the literature*. Nature Genetics, 36(7):664, 2004

[11] (2008) Humboldt-Universität zu Berlin – Ali Baba: PubMed as a graph. [Online]. Available: http://alibaba.informatik.huberlin.de/

[12] (2008) iHOP - Information Hyperlinked over Proteins. [Online].Available: http://www.ihop-net.org/UniPub/iHOP/

[13] A. Kashyap, V. Hristidis, M. Petropoulos, and S. Tavoulari: *BioNav: Effective Navigation on Query Results of Biomedical Databases*. (Short Paper), ICDE 2009, to appear. Available at http://www.cs.fiu.edu/~vagelis/publications/BioNavICDE09.pdf

[14] S. Kundu and J. Misra: *A Linear Tree Partitioning Algorithm*. SIAM J. Comput. 6(1): 151-154 (1977)

[15] W. Lee, L. Raschid, H. Sayyadi and P. Srinivasan: *Exploiting Ontology Structure and Patterns of Annotation to Mine Significant Associations between Pairs of Controlled Vocabulary Terms*. DILS 2008: 44-60

[16] D. Lindberg, B. Humphreys, and A. McCray: *The Unified Medical Language System*. Methods of Information in Medicine, 32(4):281–291, 1993.

[17] D. Maglott, J. Ostell, K.D. Pruitt and T. Tatusova: *Entrez Gene:Gene-Centered Information at NCBI*. Nucleic Acids Res. 2005 January 1; 33(Database Issue): D54–D58

[18] Medical Subject Headings (MeSH®). http://www.nlm.nih.gov/mesh/

[19] J.A. Mitchell, A.R. Aronson and J.G. Mork: *Gene Indexing: Characterization and Analysis of NLM's GeneRIFs*. In Proceedings of the AMIA Symposium, 8th–12th November, Washington, DC, pp. 460–464

[20] (2008) OMIM - Online Mendelian Inheritance in Man. [Online]. Available: http://www.ncbi.nlm.nih.gov/Omim/

[21] C. Perez-Iratxeta, P. Bork and M. A. Andrade: *Exploring MEDLINE Abstracts with XplorMed*. Drugs of Today. 2002;38:381-389

[22] C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg and U. Leser: Ali Baba: PubMed as a graph. Bioinformatics, 22(19):2444-2445, 2006

[23] (2003) PubMatrix : A Tool for Multiplex Literature Mining. [Online]. Available: http://pubmatrix.grc.nia.nih.gov/

[24] (2008) PubMed PubReMiner: A Tool for PubMed Query Building and Literature Mining. [Online]. Available: http://bioinfo.amc.uva.nl/human-genetics/pubreminer/

[25] H. Shatkay, R. Feldman: *Mining the Biomedical Literature in the Genomic Era: An Overview*. Comput. Biol. 2003;10(6):821-55

[26] (2008) Stanford University – HighWire Press. [Online]. Available: http://highwire.stanford.edu/

[27] (2008) Transinsight GmbH – GoPubMed. [Online]. Available: http://www.gopubmed.org/

[28] (2008) Vivísimo, Inc. – Clusty. [Online]. Available: http://clusty.com/

[29] (2008) XplorMed: eXploring Medline abstracts. [Online]. Available: http://www.ogic.ca/projects/xplormed/

[30] T. Zhang, R. Ramakrishnan and M. Livny: *BIRCH: An Efficient Data Clustering Method for Very Large Databases*. SIGMOD Conference 1996: 103-114

[31] Lin J, DiCuccio M, Grigoryan V, Wilbur WJ. Exploring the effectiveness of related article search in PubMed. TR, Uni Maryland, College Park. 2007 July;.

[32] Computation of Related Articles;. Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/computation.html.

[33] Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007;8(243).

[34] Synnestvedt M, Chen C. Visualizing AMIA : a medical informatics knowledge domain analysis. In: Proc AMIA Symp.; 2003. p. 1024.

[35] Synnestvedt M, Chen C, Holmes J. CiteSpace II: visualization and knowledge discovery in bibliographic databases. In: Proc AMIA Symp.; 2005. p. 724–8.

[36] Braam RR, Moed HF, van Raan AFJ. Mapping of science by combined co-citation and word analysis. I. Structural aspects. Journal of the American Society for Information Science. 1999;42(4):233–251.

[37] Tbahriti I, Chichester C, Lisacek F, Ruch P. Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE digital library. Int J Med Inform. 2006;75(6):488–495.
[38] PMC Open Access Subset;. Available from: http://www.pubmedcentral.nih.gov/about/openftlist.html.
[39] Zhiyong Lu, PubMed and beyond: a survey of web tools for searching biomedical literature, Database, Vol. 2011, Article ID baq036, doi:10.1093/database/baq036