International Journal of Advanced Computer and Mathematical Sciences
ISSN 2230-9624. Vol 3, Issue 4, 2012, pp 494-500
http://bipublication.com

A QUERY ON PUBMED RESULTS USING HIERARCHIES

Akula V Nageswara Rao and K.Eswar

Dept of CSE, B.V.C Engineering College, Odalarevu, A.P, India

[Received-09/09/2012, Accepted-01/10/2012]

ABSTRACT:

A natural way to organize biomedical citations is according to their MeSH annotations. MeSH is a comprehensive concept hierarchy used by PubMed. In this paper, we present the BioNav system, a novel search interface that enables the user to navigate large number of query results by organizing them using the MeSH concept hierarchy. First, the query results are organized into a navigation tree. At each node expansion step, BioNav reveals only a small subset of the concept nodes, selected such that the expected user navigation cost is minimized. In contrast, previous works expand the hierarchy in a predefined static manner, without navigation cost modeling.

INTRODUCTION:

Search queries on biomedical databases, such as PubMed, often return a large number of results, only a small subset of which is relevant to the user. Ranking and categorization, which can also be combined, have been proposed to alleviate this information overload problem. Result optimization and results categorization for biomedical databases is the focus of this work. The last decade has been marked by unprecedented growth in boththe production of biomedical data and the amount of published literature discussing it. The MEDLINE database, on which the PubMed search engine operates, contains over 18 million citations and is currently growing at the rate of 500,000 new citations each year [31].

Literature survey:

literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things r satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into account for developing the proposed system. Many solutions have been proposed to address this problem—commonly referred to as information overload [1], [2], [3] ,[9], [16]. These approaches can be broadly classified into two classes: ranking and categorization—which can also be combined. Ranking presents the user with a list of results ordered by some metric of relevance [9] or by content similarity to a result or a set of results [16]. In categorization [1], [2], [3], query results are grouped based on hierarchies, keywords, tags, or attribute values. User studies have

demonstrated the usefulness of categorization in finding relevant results of exploratory queries [12]. While ranked results are useful when the ranking function is aligned with user preferences or the result list is small in size, categorization is generally employed by users when ranking fails or the query is too "broad" [12]. BioNav belongs primarily to the categorization class, which is especially suitable for this domain given the rich concept hierarchies (e.g., MeSH [19]) available for biomedical data. An intuitive way to categorize the results of a query on PubMed is by using the MeSH static concept hierarchy [19], 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: 1) by being explicitly annotated with them, and 2) by mentioning those in their text (see Section 7 for details). Since these associations are provided by PubMed, 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. Fig. 1 displays a snapshot of such an interface where shown next to each node label is the count of distinct citations in the sub-tree rooted at that node. A typical navigation starts by revealing the children of the root ranked by their citation count, and is continued by the user expanding on or more of them, revealing their ranked children and so on, until she clicks on a concept and inspects the citations attached to it. A similar interface and navigation method is used by e-commerce sites, such as Amazon and eBay. For this example interaction, we assume that some of the citations the user is interested in are available on the three indicated concepts corresponding to three independent lines of research related to prothymosin, and therefore the user is interested in navigating to these concepts. These include, "Histones," which play a role in gene regulation and are essential for virus replication and tumor growth, "Cell Growth Processes" and "Transcription, Genetic," a key process for synthesis and replication of RNA and thus plays an important role in the duplication of cancer cells.

EXISTING SYSTEM:

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. Other biological sources, such as Entrees Gene and OMIM, witness similar growth. As claimed in previous work, 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.

IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Main Modules:-

Modules

1. Query Search process module (or) Biomedical Search Systems module

PubMed– using a keyword search interface. Currently, in an exploratory scenario where the user tries to find citations relevant to her line of research and hence not known a priori, she 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 she has over-specified the query, in which case relevant citations might be excluded from the final query result.

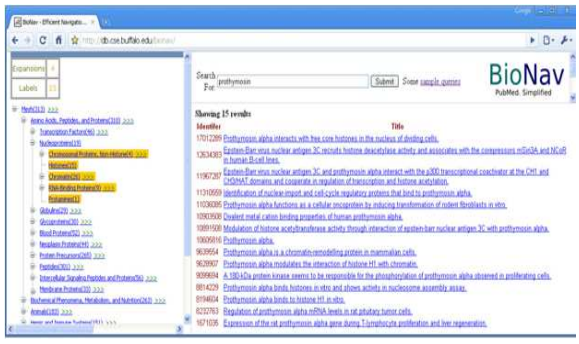
Query on PubMed is using the MeSH static concept hierarchy , thus utilizing the initiative of the US National Library of Medicine (NLM) to build and maintain such a comprehensive structure. Each citation in MEDLINE is associated with several MeSH concepts in two ways: (i) by being explicitly annotated with them, and (ii) by mentioning those in their text . Since these associations are provided by PubMed, 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

2.Dynamic navigation tree module

navigation tree. Figdisplays a snapshot of such an interface where shown next to each node label is the count of distinct citations in the subtree rooted at that node. A typical navigation starts by revealing the children of the root ranked by their citation count, and is continued by the user expanding on or more of them, revealing their ranked children and so on, until she clicks on a concept and inspects the citations attached to it. A similar interface and navigation method is used by e-commerce sites, such as Amazon and eBay. For this example, we assume that the user will navigate to the three indicated concepts corresponding to three independent lines of research related to prothymosin

BioNav introduces a dynamic navigation method that depends on the particular query result at hand and is demonstrated in Fig The query results are attached to the corresponding MeSH concept nodes as in Fig. but then the navigation proceeds differently. The key action on the interface is the expansion of a node that selectively reveals a ranked list of descendant (not necessarily children) concepts, instead of simply showing all its children.

3.Hierarchy navigation web (interface) search module



BioNav belongs primarily to the categorization class, which is ideal for this domain given the rich concept hierarchies (e.g., MeSH) 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.

4. Query Workload online operation module

On-Line Operation. Upon receiving a keyword query from the user, BioNav executes the same query against the MEDLINE database and retrieves only the IDs (Pub Med Identifiers) of the citations in the query result. This is done using the ESearch utility of the Entrez Programming Utilities (eUtils) . eUtils are a collection of web interfaces to PubMed for issuing a query and downloading the results with various levels of detail and in a variety of formats. Next, the navigation tree is constructed by retrieving the MeSH concepts associated with each citation in the query result from the BioNav database. This is possible since MeSH concepts have tree identifiers encoding their location in the MeSH hierarchy, which are also retrieved from the BioNav database. This process is done once for each user query.

PROPOSED SYSTEM:

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. These approaches can be broadly classified into two classes: ranking and categorization, which can also be combined.

An intuitive way to categorize the results of a query on Pub Med is using the Mesh static concept hierarchy, 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:

- by being explicitly annotated with them,
- By mentioning those in their text (see Section 7 for details). 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, Bionic 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.

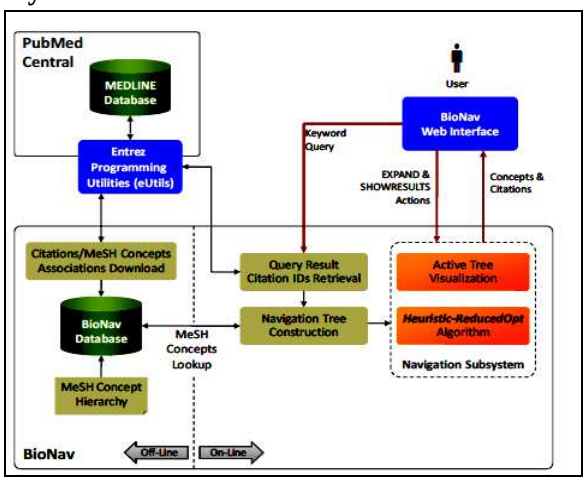
TOPDOWN Cost Model:

The cost model, which is inspired by a previous work, 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.

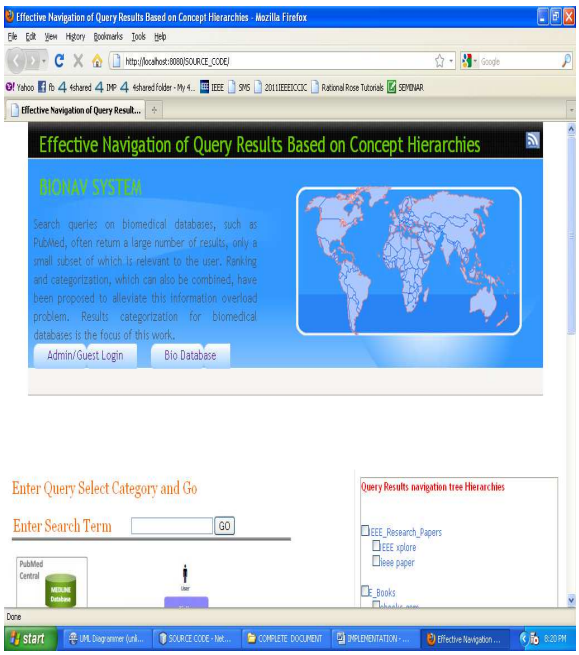
System Architecture:

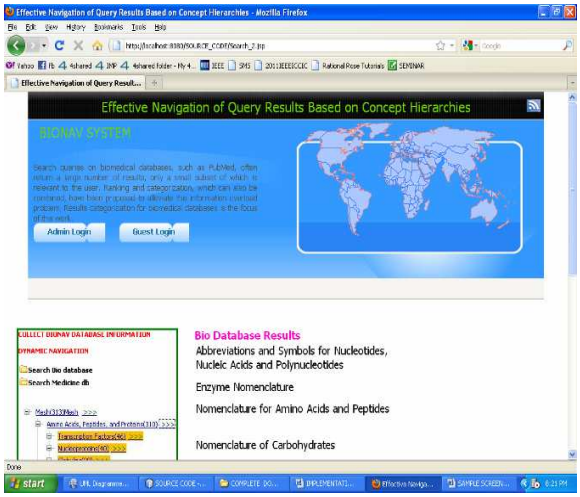


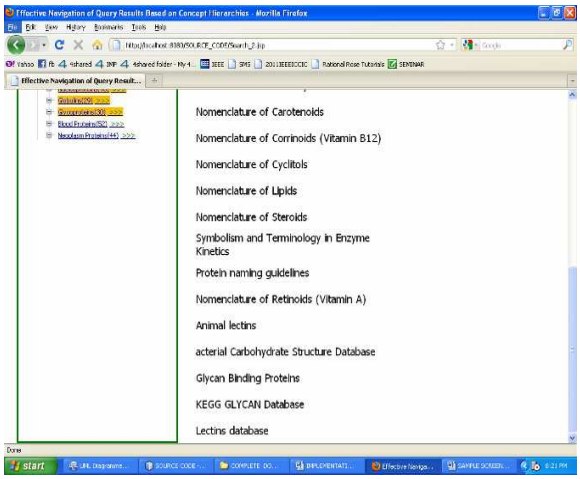
Algorithm used:

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

Sample screen shots :







SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

CONCLUSION

Information overload is a major problem when searching biomedical databases such as PubMed, where typically a large number of citations are returned, of which only a small subset is relevant to the user. 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. Our complexity result proved that the problem of expanding the navigation tree in a way that minimizes the user's navigation cost is NP-complete. A feasible (for small trees) optimal algorithm and an efficient heuristic were developed. Experimental results validated the effectiveness of the proposed heuristic for diverse sets of queries and navigation trees, when compared to categorization systems using a static navigation method.

REFERENCES

[1] J S. Agrawal, S. Chaudhuri, G. Das and A. Gionis: Automated Ranking of Database Query Results. In Proceedings of First Biennial 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 Infinite Expansions, 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] Medical Subject Headings (MeSH)(2010), [Online] Available: http: //www.nlm.nih.gov/mesh/