Optimizing Query results using Middle Layers Based on Concept Hierarchies

G.Kumari [€],I.Gayathri Devi [*]
G.Kumari Asst.Professor, CSE Department
I.Gayathri Devi Asst.Professor, CSE Department
Pragati Engineering College,Surampalem,Kakinada,India

Abstract

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 als o 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. 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 BioIntelR (BIR) system, adopts the BioNav system enables the user to navigate large number of query results by organizing them using the MeSH concept hierarchy. First, BioIntelR (BIR) system prompts the user for the search criteria and the system automatically connects to a middle layer created at the application level which directs the query to the proper valid query path to select correct criteria of the search result from the biomedical database. The query results are organized into a navigation tree. At each node expansion step, BIR system reveals only a small subset of the concept nodes, selected such that the expected user navigation cost is minimized. In contrast, to the previous systems, the BIR system outperforms and optimizes the query result time and minimizes query result set for easy user navigation, Data Warehousing.

***Keywords*—Interactive data exploration and discovery, search process, graphical user interfaces, interaction styles,BioNav System,MESH.**

1. Introduction

The last decade has been marked by unprecedented growth in both the 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 [20]. Other biological sources, such as Entrez Gene [18] and OMIM [21], witness similar growth. Biologists, chemists, medical and health scientists are used to 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 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 resubmits 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.

A typical navigation starts with revealing the children of the root ranked by their citation count, and is continued with expanding one or more of them, revealing their ranked children and so on. Further, the user may click on a concept and inspect the attached citations. A similar interface and navigation method is used by GoPubMed [5] and e-commerce sites, such as Amazon and eBay.

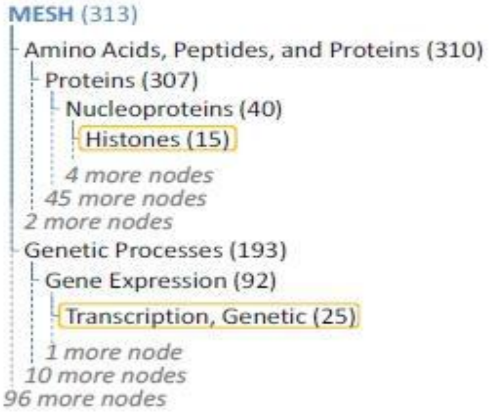
MESH (313)
Amino Acids, Peptides, and Proteins (310)
Proteins (307)
Nucleoproteins (40)
Histones (15)
4 more nodes
45 more nodes
2 more nodes
Genetic Processes (193)
Gene Expression (92)
Transcription, Genetic (25)
1 more node
10 more nodes
96 more nodes

Figure 1. Static Navigation on the Mesh Concept Hierarchy

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 resubmits 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.

Figure 1 displays 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. For this example, we assume that the user queries MEDLINE for the nucleoprotein ―prothymosin‖ and his personal interests are reflected in the two indicated concepts, corresponding to two independent lines of research related to prothymosin. A typical navigation starts with revealing the children of the root ranked by their citation count, and is continued with expanding one or more of them, revealing their ranked children and so on. Further, the user may click on a concept and inspect the attached citations. A similar interface and navigation method is used by GoPubMed [5] and e-commerce sites, such as Amazon and eBay.

The static navigation method −same for every query result−is problematic when the MeSH hierarchy (or one with similar properties) is used for categorization for the following reasons: 1. The massive size of the MeSH hierarchy (with 48,441 concept nodes) makes it challenging for the users to effectively navigate to the desired concepts and browse the associated citations.

2. A substantial number of duplicate citations are introduced in the navigation tree of Figure 1, since each one of the 313 distinct citations is associated with several concepts. Specifically, the total count of citations in Figure 1 is 40,195.

So, The BIR system is developed to facilitate the keyword search on PubMed to using MeSH concept hierarchy. The Proposed system, accepts the user search criteria and specifies the valid query path by using middle layer constructed at the application level before running on the Bio Medical databases by incorporating the BioNav methods to get the meaningful results which leads to the further data analysis.

This paper is organized as follows: Section 2 describes related work, Section 3 describes the Proposed System architecture Work nature, Section 4 presents experiment Evolution. Section 5 presents conclusion and future scope

2. Related Work

Several systems have been developed to facilitate keyword search on PubMed using the MeSH concept hierarchy. Pubmed itself allows the user to search for citations based on MeSH annotations. A keyword query "histones [MeSH Terms]" will retrieve all citations annotated with the MeSH term "histones" in the MeSH hierarchy. The user can also limit her search to a MeSH term by using additional filters, e.g., "[majr]" to filter out all citations in the query result that don't have the term as their major term. These filters can be combined by using the Boolean connectives AND, OR, and NOT. This interface poses significant challenges, even to experienced users, since the annotation process is manual and thus prone to errors. The closest to BioNav is GoPubMed which implements a static navigation method on the results of PubMed. GoPubMed lists a predefined list of high-level MeSH concepts, such as "Chemicals and Drugs," "Biological Sciences," and so on, and for each one of them displays the top-10 concepts. After a node expansion, its children are revealed and ranked by the number of their attached citations, whereas BioNav reveals a selective and dynamic list of descendant (not always children) nodes ranked by their estimated relevance to the user's query. Further, BioNav uses a cost model to decide which concepts to display at each step.

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

when the MeSH hierarchy (or one with similar properties) is used for categorization for the following reasons

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. for example, shows the initial expansion of the root node where only eight (highlighted) descendants are revealed compared to 98 children shown in Fig. The concepts are ranked by their relevance to the user query and the number of them revealed depends on the characteristics of the query results. Next, assuming the user is interested in the "Amino Acids ..." node and judging that the 310 attached citations is still a big number, she expands it by clicking on the ">> >" hyperlink next to it in Fig. The user inspects the six concepts revealed and decides that she is not interested in any of them. Hence, she expands the "Amino Acids ..." node one more time in Fig.revealing four additional concepts. Note that "Nucleoproteins" is an example of a descendant node being revealed, since its parent node "Proteins" is not revealed in Fig. the user expands the "Nucleoproteins" node and reveals "Histones," one of the three key concepts for the query. In the last step of the interaction, the user clicks on the "Histones" hyperlink and the 15 corresponding citations are displayed in a separate frame as shown in Fig. To reach "Histones" using the BioNav navigation method, only 23 concepts are revealed, after four node expansions, compared to 152 concepts, also after four expansions, with the static navigation method of Fig. For each expansion, the displayed descendant concepts are chosen in a way that the expected navigation cost is minimized, based on an intuitive navigation cost model.

In addition to the static hierarchy navigation works mentioned above, there are works on dynamic categorization of query results (e.g., the Clusty search engine)which create unsupervised query-dependent results clusters, but do not study how the Clusters should be navigated. BioNav is distinct since it offers dynamic navigation on a predefined hierarchy, as is the MeSH concept hierarchy. Another difference is that BioNav uses a navigation cost model to minimize the navigation cost framework and BioIntelR

3. Overview

Information overload is a common phenomenon encountered by users searching biomedical databases such as PubMed. We encounter this problem, we resolve this problem by optimizing the query result time and minimize query result set for easy user navigation.

3.1 Architecture of BioIntelR System

The propose BIR system consists combination of :

1. Web interfaces
2. Middle layer
3. Navigation system,
4. Programming utilizes
5. Data base

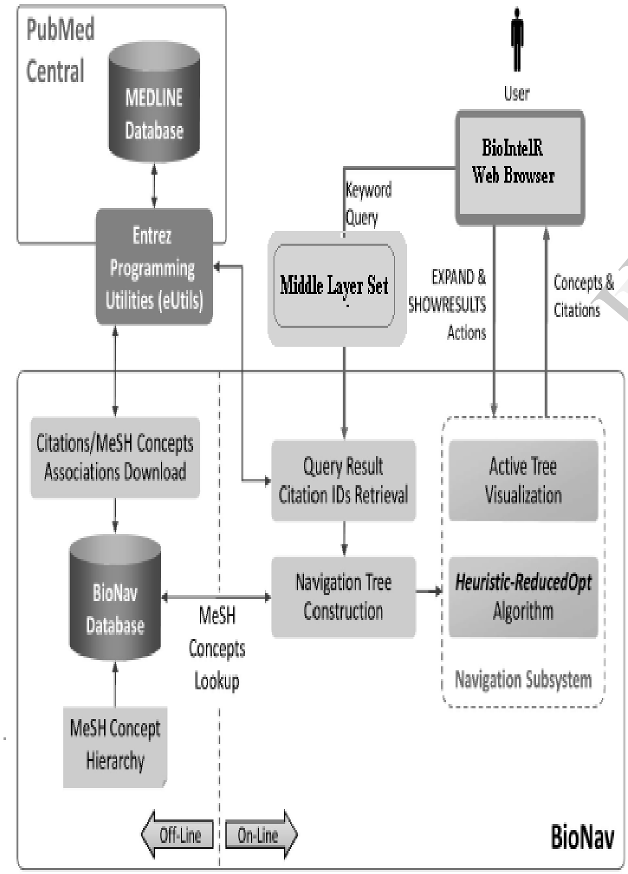
PubMed Central
MEDLINE Database
Entrez Programming Utilities (eUtils)
User
BioIntelR Web Browser
Keyword Query
Middle Layer Set
EXPAND & SHOWRESULTS Actions
Concepts & Citations
Citations/MeSH Concepts Associations Download
Query Result Citation IDs Retrieval
Active Tree Visualization
BioNav Database
MeSH Concepts Lookup
Navigation Tree Construction
Heuristic-ReducedOpt Algorithm
Navigation Subsystem
MeSH Concept Hierarchy
Off-Line
On-Line
BioNav

Figure 2. BioIntelR system architecture

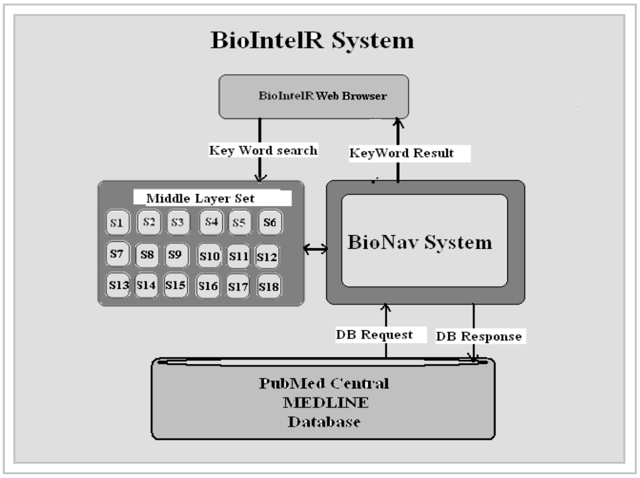

Figure 3. BioIntelR System

Upon receiving a keyword query from the user, BioIntelR sends the query and Visualize the query results, BioNav executes the same query against the MEDLINE database and retrieves only the IDs (PubMed Identifiers) of the citations in the query result. The user interacts with system by using BioIntellR web browser to find the effective results of the search criteria from PubMed. Previously the BioNav system, once the user issues a keyword query, PubMed—BioNav uses the Entrez Programming Utilities (eUtils) [7]—returns a list of citations, each associated with several MeSH concepts. BioNav constructs an Initial Navigation Tree by attaching to each concept node of the MeSH concept hierarchy a list of its associated citations BioNav reduces the size of the initial navigation tree by removing the nodes with empty results lists, while preserving the ancestor/descendant relationships.

Middle Layer: The role of the middle is to provide an easy to use and understand interface for user to search criteria against database to get the minimal result set for easy navigation and it reduces search result time.

The middle layer is a file mainly consists of the schema of objects is created according underlying database, the file contains connection parameter to connect the database, the middle layer Maps the search keywords to the database and validated path for the search criteria .the layer acts as bridge between the user interface and the database. The schemas that we created must be relevant to the end user business environment and vocabulary.

Navigation system . 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 subtree rooted at concept node n: EXPAND, SHOWRESULTS, IGNORE, BACKTRACK this navigation process continues until the user finds all the citations she is interested in.

3.2 Work Nature of the Proposed System

The main aim of the proposed system is to search effective results from millions citations.

The proposed system BioIntelR (BIR)(contains set of Middle layers and any Bio medical search tool we consider the BioNav System) accepts the user search key words and prompts the user for specific filter fields, the system accepts user request, and choose the effective middle layer from the middle layer set to effectively process request .The layer acts as a bridge between the user interface and BioIntelR system.

Offline Preprocessing. The BioNav database is first populated with the MeSH hierarchy, which is available online [19] and has more than 48,000 concept nodes. Then, the BioNav database is populated with the associations of the MEDLINE citations to MeSH concepts. These associations are not directly provided by the Entrez Programming Utilities (eUtils), so we had to implement the following method to infer these associations. For each concept in the MeSH hierarchy, we issued a query on PubMed using the concept as the keyword. For each citation ID in the query result, we added to a table in the BioNav database the tuple <concept; citationID>. Alternatively, we could determine the associations by using the MeSH concepts that each citation is annotated within the MEDLINE database. This information is available through eUtils.In this case though, the navigation trees of BioNav would not be very informative, since each citation is annotated with 20 concepts on average in MEDLINE, while the PubMed indexing associates each citation with approximately 90 concepts on average (and include the 20 from MEDLINE.) Given the number of concepts in the MeSH hierarchy, the number of citations in MEDLINE (_18 million), and the PubMed eUtils restrictions on the number of queries that can be executed within a certain period of time, it took almost 20 days to collect all the <concept; citationID> tuples. In the end, there were almost 747 million such tuples. To improve the selection queries on this table, we denormalized it by concatenating all concepts associated with each citation into a comma-separated list, that is <citationID; ðconcept1; concept2; . . .Þ>:In this work, we assume the data set D to be fixed. However, in practice, D changes frequently as new citations are added and existing citations are updated to include new terms from the MeSH hierarchy. In this case, we assume that D is refreshed periodically by an offline process that issues queries to PubMed using the concept keyword and updates the concept counts and rows of retrieved citations.

4. Experimental Evaluation

We evaluated the difference between the BioIntelR and BioNav systems in terms of both average Navigation cost and expansion time performance Other traditional measures of quality, such as precision and recall, are not applicable to our scenario as the objective is to minimize the tree navigation cost and not to classify

we show that the BioIntelR method, which is evaluated using middle layer and adopted BioNav system and the BioNav system Heuristic-ReducedOpt algorithm, leads to considerably smaller navigation cost for a set of real queries on the MEDLINE database and navigations on the MeSH hierarchy. we compare the optimal algorithm (Opt-EdgeCut) with Heuristic-ReducedOpt and show that the heuristic is a good approximation of the optimal. These experiments were executed on a reduced navigation tree (_20 nodes), constructed from the original query navigation tree for each query, The experiments were executed on a Windows XP Professional machine with 3 GHz CPU and 2 GB of main memory, running Windows XP Professional. All algorithms were implemented in Java and Oracle 10g was used as the database.

Navigation Cost Evaluation. Figure .5 compares the Overall navigation cost of BioIntelR over BioNav for the biochemistry query set only. BioIntelR performs better than BioNav for all queries.

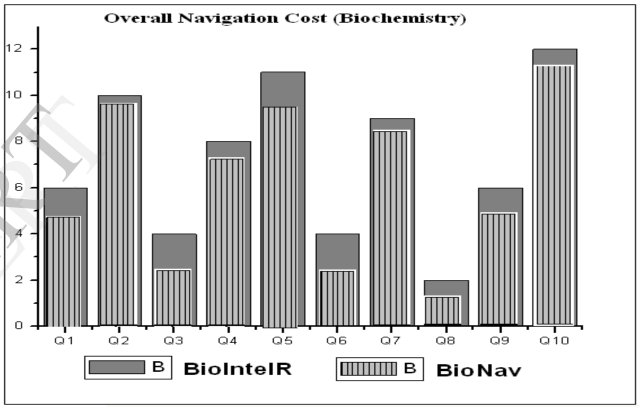

Figure 4. Overall Navigation cost comparision

5. Conclusion and Future work

Information overload is a common phenomenon encountered by users searching biomedical databases such as PubMed. We address this problem by organizing the query results according to their associations to concepts of the MeSH concept hierarchy and propose a dynamic navigation method on the resulting navigation tree. Each node expansion on the navigation tree, reveals a small set of nodes, selected from among its descendents, and the nodes are selected such that the information overload observed by the user is minimized. We formally stated the underlying framework and the navigation and cost models used for evaluation of our approach. We prove that the problem of selecting the set of nodes that minimize the navigation cost is NP-complete, we propose an efficient heuristic

6. References

[1] IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 4, april 2011

[2]J.S.Agrawal,S.Chaudhuri,G.Das,andA.Gionis“Automated Ranking of Database Query Results,” Proc. First Biennial Conf.Innovative Data Systems Research, 2003.

[3] K. Chakrabarti, S. Chaudhuri, and S.W. Hwang, “AutomaticCategorization of Query Results,” Proc. ACM SIGMOD, pp. 755-766, 2004.

[4] Z. Chen and T. Li, “Addressing Diverse User Preferences in SQLQuery-Result Navigation,” Proc. ACM SIGMOD, pp. 641-652,2007.

[5] L. Comtet, Advanced Combinatorics: The Art of Finite and InfiniteExpansions, pp. 176-177, Reidel, 1974.

[6] R. Delfs, A. Doms, A. Kozlenkov, and M. Schroeder, “GoPubMed:Ontology-Based Literature Search Applied to Gene Ontology and PubMed,” Proc. German Conf. Bioinformatics, pp. 169-178, 2004.

[7] D. Demner-Fushman and J. Lin, “Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering,” Proc. Int’l Conf. Computational Linguistics and Ann.Meeting of the Assoc. for Computational Linguistics, pp. 841-848, 2006.

[8] Entrez Programming Utilities, http://www.ncbi.nlm.nih.gov/ entrez/query/static/eutils_help.html, 2008.

[9] U. Feige, D. Peleg, and G. Kortsarz, “The Dense k-Subgraph Problem,” Algorithmica, vol. 29, pp. 410-421, 2001.

[10] V. Hristidis and Y. Papakonstantinou, “DISCOVER: Keyword Search in Relational Databases,” Proc. Int’l Conf. Very Large Data Bases (VLDB), 2002.

[11] R. Hoffman and A. Valencia, “A Gene Network for Navigating the Literature,” Nature Genetics, vol. 36, no. 7, p. 664, 2004.

[12] iHOP—Information Hyperlinked over Protein, http://www. ihop-net.org/UniPub/iHOP/, 2008.

[13] M. Kaki, “Findex: Search Results Categories Help When Document Ranking Fails,” Proc. ACM SIGCHI Conf. Human Factors in Computing Systems, pp. 131-140, 2005.

[14] A. Kashyap, V. Hristidis, M. Petropoulos, and S. Tavoulari, “BioNav: Effective Navigation on Query Results of Biomedical Databases,” Proc. IEEE Int’l Conf. Data Eng. (ICDE), (short paper), pp. 1287-1290, 2009.

[15] S. Kundu and J. Misra, “A Linear Tree Partitioning Algorithm,” SIAM J. Computing, vol. 6, no. 1, pp. 151-154, 1977.