International Journal of Engineering Research and Development
ISSN: 2278-067X, Volume 2, Issue 1 (July 2012), PP. 01-06
www.ijerd.com

Efficient Navigation of Query Results Using Concept Hierarchies

Shilpa Mentada[1], B. Govinda Lakshmi[2]
[1] 2/2 M.TECH SE, Department of CSE, Sri Sivani College of Engineering, Chilakapalem, Srikakulam, AP, India
[2]Associate Professor, Department of CSE, Sri Sivani College of Engineering, Chilakapalem, Srikakulam, AP, India

Abstract*—with the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web data research has encountered a lot of challenges. To overcome this so many researchers proposed various techniques but still there is a requirement. To fulfill this, in this paper we demonstrate the BioNav system, a novel search interface for biomedical databases, such as PubMed. BioNav enables users to navigate large number of query results by categorizing them using MeSH; a comprehensive concept hierarchy used by PubMed. Once the query results are organized into a navigation tree, BioNav reveals only a small subset of the concept nodes at each step, selected such that the expected user navigation cost is minimized.

***Keywords*—Web Mining, World wide web, Web Personalization, BioNav System, MeSH;**

I. 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 the database is growing at the rate of 500,000 new citations each year [1-3]. Keyword search queries on these databases return a large results set from which only a small portion is relevant for the user. Many solutions have been proposed to address this problem – commonly referred to as information-overload [3-6]. 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 available for biomedical data, such as MeSH. Each citation in MEDLINE is associated with several MeSH concepts in two ways: (i) by being explicitly annotated with them, and (ii) by mentioning them 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 concept hierarchy

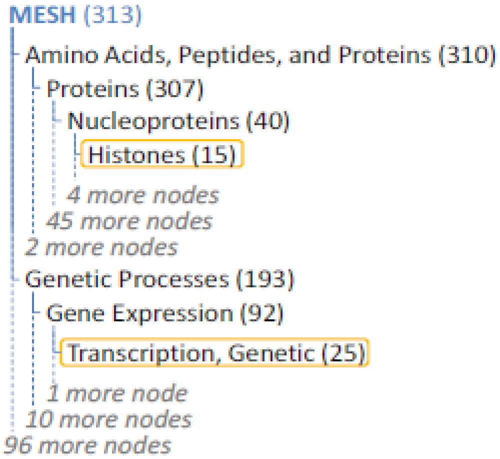

Figure 1. Static Navigation on the MeSH Concept Hierarchy

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 above *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:

- 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.
- A substantial number of *duplicate* citations are introduced in the navigation tree of Figure 1, since each one of the 313 d*istinct c*itations is associated with several concepts. Specifically, the total count of citations in Figure 1 is 40,195.

II. BACKGROUND

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 [7-11], 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.

The user then navigates this tree structure, in a top-down fashion, exploring the concepts of interest while ignoring the rest Node that the user is not aware that the relevant results are available specifically on these nodes—she is only interested in narrowing down the results, using a familiar concept hierarchy, instead of examining all the results The above static—same for every query result—navigation method is problematic when the MeSH hierarchy (or one with similar properties) is used for categorization for the following reasons.

III. PROPOSED SYSTEM ARCHITECTURE

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.

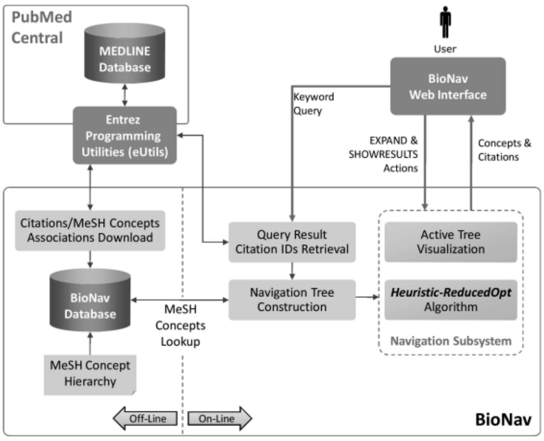

Fig. 2: BioNav System Architecture

A. Architecture of BioIntelR System

The propose BIR system consists combination of (shown in Fig 2 and 3 which shows the actual flow):

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

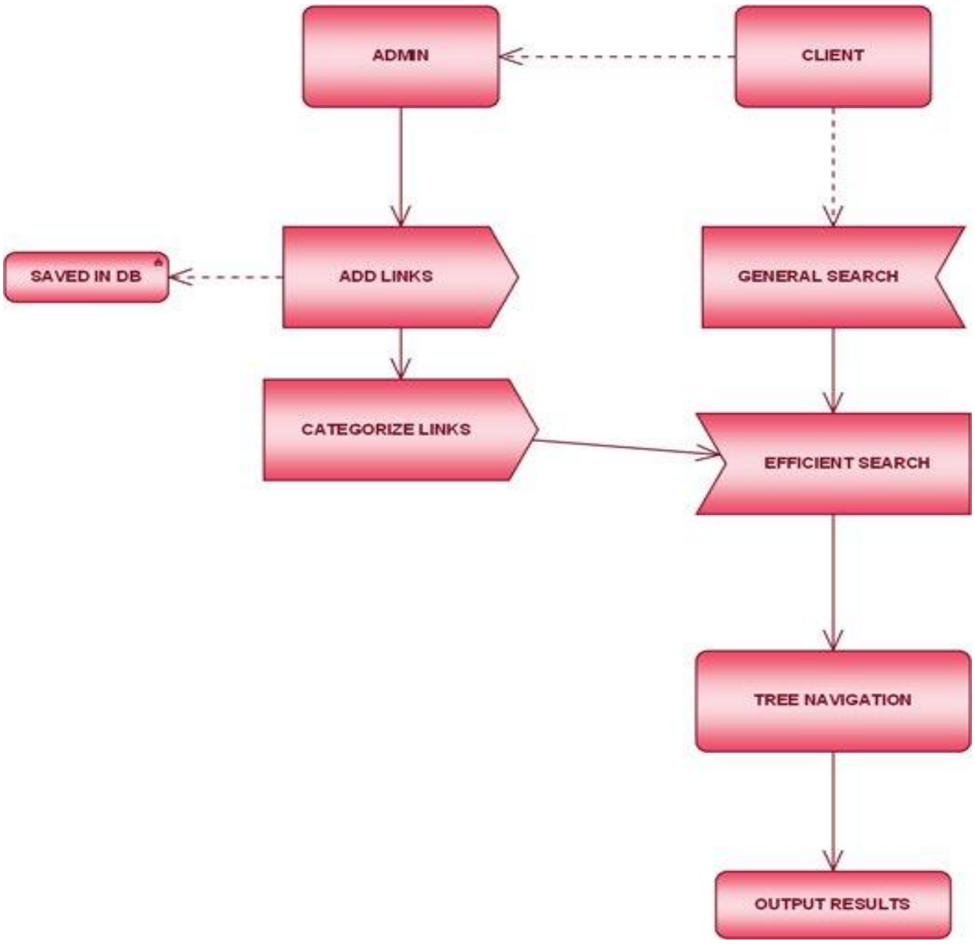

Fig 3 system flow

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 BioIntelR web browse 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—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. The MeSH concept hierarchy is the starting point of the framework and is defined as follows: Concept Hierarchy, Navigation Tree, Valid EdgeCut, Active Tree, Active Tree Visualization, and The navigation model of BioNav which is used to device and evaluate algorithms .

Web Interface: Web interface is the user interface interacts with the BiointelR system, by specifying the search criteria by specifying the search key words to visualize the optimized results from the system.

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.

Programming Utilities: Entrez Programming Utilities returns a list of citations, each associated with several MeSH concepts

Data Bases: MEDLINE database, fig .3 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. The BioNav database is first populated with the MeSH hierarchy, which is available online [12-15] and has more than 48,000 concept nodes. Then, the BioNav database is populated with the associations of the MEDLINE citations to MeSH concepts.

B. Algorithm

```
Procedure. GenReducedTree
Input: Initial Navigation Tree I(n), the target concept c, and the desired number max N of nodes in the reduced tree
Output: A reduced tree with at most maxN nodes,including c
    1.  collect all nodes of I(n) in list L
    2.  create list $L^1$ to store the nodes of the reduced tree
    3.  add to $L^1$ a concept node in L with the same label as c and all its ancestors
    4.  while (sizeof(L)$\leq$ maxN) repeat
    5.  select a node c0 uniformly at random from L
    6.  add $c^1$ and all its ancestors to $L^1$, excluding duplicates
    7.  create a tree $I^1$(n) from the nodes in $L^1$ preserving the parent-child relationship
    8.  return $I^1$(n)
```

IV. IMPLEMENTATION

In this paper our proposed method consists following modules.

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

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.

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.

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.

V. RESULT ANALYSIS

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, since Opt-EdgeCut is prohibitively expensive for most navigation trees. Finally, shows that the execution time of Heuristic-ReducedOpt is small enough to facilitate interactivetime use navigation. 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. These comparisons shown in Fig 4 and 5.

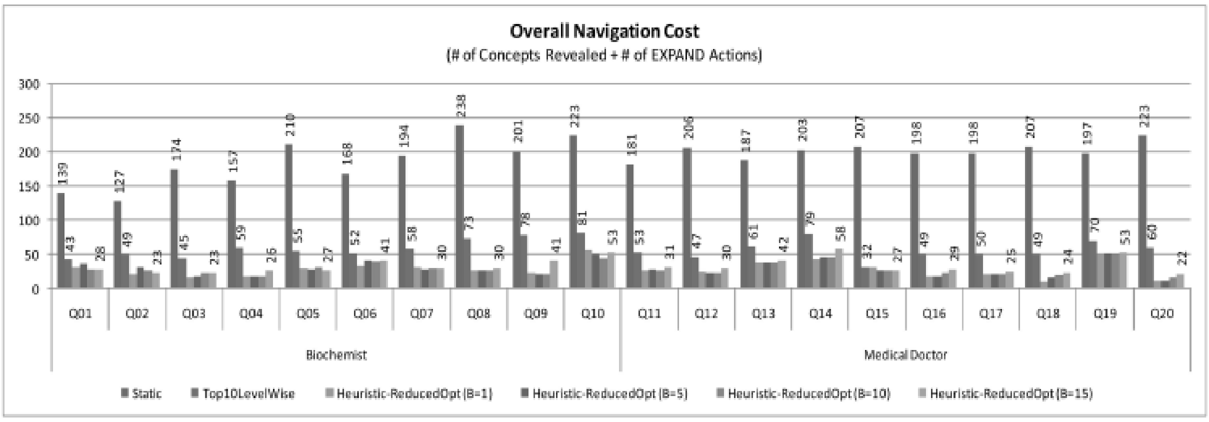

Fig. 4. Overall navigation cost comparison for biochemistry and medicine

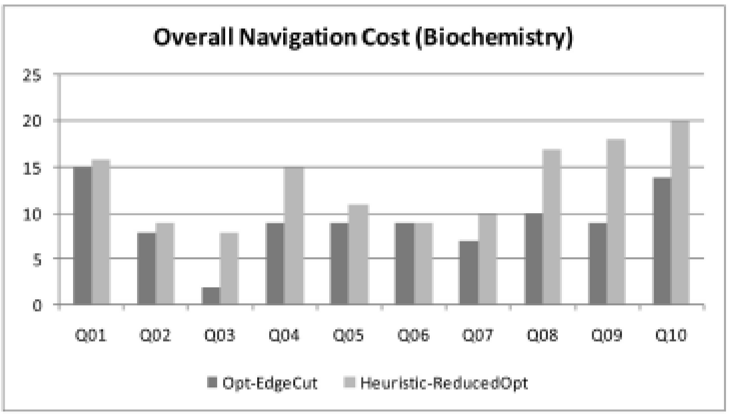

Fig. 5. Overall navigation cost comparison

VI. CONCLUSIONS

In this paper, we presented the BioNav system which organize the query results according to user 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.

REFERENCES

[1]. 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, 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 SQL-Query-Result Navigation. SIGMOD Conference 2007: 641-652

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

[5]. Transinsight GmbH – GoPubMed. http://gopubmed.org

[6]. Stanford Univ. (2008),"HighWire Press", [Online] Available: http://highwire.stanford.edu/

[7]. J.S. Agrawal, S. Chaudhuri, G. Das, and A. Gionis, "Automated Ranking of Database Query Results," Proc. First Biennial Conf. Innovative Data Systems Research, 2003.

[8]. K. Chakrabarti, S. Chaudhuri, and S.W. Hwang, "Automatic Categorization of Query Results," Proc. ACM SIGMOD, pp. 755-766, 2004.

[9]. D. Demner-Fushman, 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.

[10]. 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," Proc. Fifth Int'l Workshop Data Integration in the Life Sciences (DILS), pp. 44-60, 2008.

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

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

[13]. A. Kashyap, V. Hristidis, M. Petropoulos, 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.

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

[15]. V. Hristidis, Y. Papakonstantinou,"DISCOVER: Keyword Search in Relational Databases", Proc. Int'l Conf. Very Large Data Bases (VLDB), 2002.