

BioNav: Effective Navigation on Query Results of Biomedical Databases

Abhijith Kashyap^{*1}, Vagelis Hristidis[#], Michalis Petropoulos^{*2}, Sotiria Tavoulari⁺

^{*}*Department of Computer Science and Engineering, SUNY at Buffalo
201 Bell Hall, Buffalo, NY 14260-2000, USA*

¹rk39@cse.buffalo.edu

²mpetropo@cse.buffalo.edu

[#]*School of Computing and Information Sciences, Florida International University
11200 S.W. 8th Street, Miami, FL 33199, USA*

vagelis@cis.fiu.edu

⁺*Department of Pharmacology, Yale University*

333 Cedar Street, New Haven, CT 06520-8066, USA

sotiria.tavoulari@yale.edu

Abstract— Search queries on biomedical databases like 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. 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 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 modelling. We show that the problem of selecting the best concepts to reveal at each node expansion is NP-complete and propose an efficient heuristic as well as a feasible optimal algorithm for relatively small trees. We show experimentally that BioNav outperforms state-of-the-art categorization systems with respect to the user navigation cost. We have implemented BioNav for the MEDLINE database at <http://db.cse.buffalo.edu/bionav>.

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 currently growing at the rate of 500,000 new citations each year [14]. So do other biological sources like Entrez Gene [12] and OMIM [15]. As claimed in previous work [19], the ability to rapidly survey this literature constitutes a necessary step toward both the design and the interpretation of any large scale experiment. 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 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.

As an example, 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]-[3], [8]. These approaches can be broadly classified into two categories; ranking and categorization, which can also be combined.

The primary focus of BioNav is on categorization techniques, which are ideal given the rich concept hierarchies (e.g., MeSH [13]) 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 PubMed is using the MeSH static concept hierarchy [13], 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 (see Section VII 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 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, like Amazon and eBay. For this example, we assume that the user will navigate to the four indicated concepts corresponding to four independent lines of research related to prothymosin.

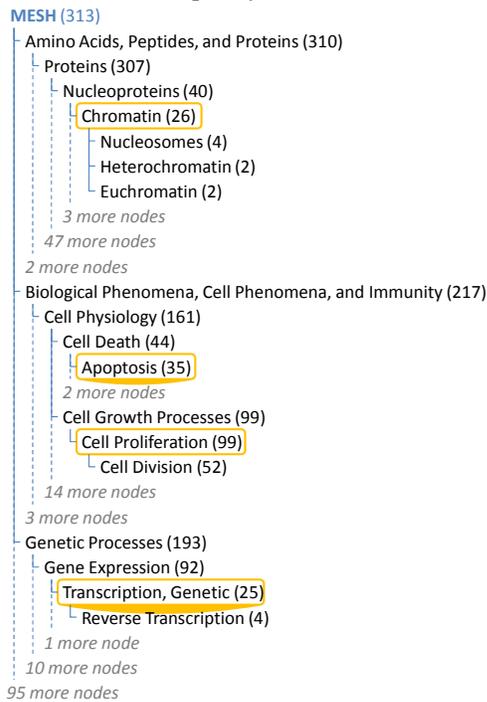


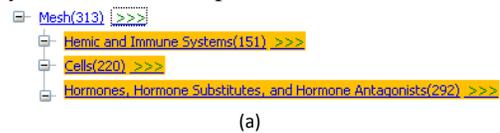
Fig. 1 Static Navigation on the MeSH Concept Hierarchy¹

The above *static* –same for every query result– navigation method is problematic when the MeSH hierarchy is used for categorization for the following reasons:

- The massive size of the MeSH hierarchy (over 48,000 concept nodes) makes it challenging for the users to effectively navigate to the desired concepts and browse the associated records. Even if we remove from the MeSH concept nodes with no citations attached to them, the 313 distinct query results for “prothymosin” are attached to 3,940 nodes, which is the actual size of the navigation tree in Fig. 1. Combined with the fact that the MeSH hierarchy is quite bushy on the upper levels, this means that the user has to inspect, for example, a total of 123 concept nodes before she reaches the indicated concepts “Cell Proliferation” and “Apoptosis”; a number

comparable to the distinct citation count in the query result. A common practice [21] for hierarchy navigation is to show only a subset of a node’s children, since there is typically a dominant concept count. Unfortunately, this is not the case for the 98 children of the root in Fig. 1, for example, as evidenced by the three comparable citation counts shown (310, 217 and 193).

- A substantial number of *duplicate* citations are introduced in the navigation tree of Fig. 1, since each one of the 313 distinct citations is associated with several concepts. Specifically, the total count of citations in Fig. 1 is 30,895. Naturally, the user would like to know which concepts fragment the query result into subsets of citations with as few duplicate citations as possible across them. Currently, the only way to figure this out using the interface in Fig. 1 is to click on different concept nodes and inspect the attached citations. As an example, the query results for “prothymosin” are related to four independent lines of research, represented by the four indicated concepts in Fig. 1, which are hard to locate. Among the total 185 citations attached to the four indicated concept nodes, only 38 of them are duplicates.



(a)



(b)



(c)

Fig. 2 BioNav Dynamic Navigation

BioNav introduces a *dynamic* navigation method that depends on the particular query result at hand and is demonstrated in Fig. 2. The query results are attached to the corresponding MeSH concept nodes as in Fig. 1, but then the

¹ The complete tree can be seen at <http://db.cse.buffalo.edu/allnav>

navigation proceeds differently. The key action on the interface is the *expansion* of a node that selectively reveals a ranked list of descendent (not necessarily children) concepts, instead of simply showing all its children.

Fig. 2a, for example, shows the initial expansion of the root node where only 3 (highlighted) descendants are revealed compared to 98 children shown in Fig. 1. The number of revealed concepts depends on the characteristics of the query results and the fine tuning of the cost model as discussed in Section III. The user inspects the revealed concepts and decides to expand the root two more times, as shown in Fig. 2b, revealing a total of 11 descendants. The concepts are ranked by their relevance to the user query. Next, assuming the user is interested in the “Biological Phenomena...” node and judging that the 217 attached citations is still a big number, she expands it by clicking on the “>>>” hyperlink next to it in Fig. 2c. Only the 5 highlighted descendent nodes, out of the 25 total descendent nodes, are revealed. Note that the “Cell Growth Processes” node, which is a child of “Biological Phenomena...” and parent of “Cell Proliferation” (one of the four key concepts for the query), is not shown. Instead, “Cell Proliferation” is immediately revealed since it has the same 99 citations as “Cell Growth Processes” and it is more specific. “Apoptosis” can be revealed in the next navigation step by expanding “Cell Death”.

For each expansion, the displayed descendent concepts are chosen in a way that the expected navigation cost is minimized, based on an intuitive navigation cost model we present. The cost model estimates the exploration probability for a node based on its selectivity, that is, the ratio of attached citations before and after the query. The navigation cost for a concept node is also proportional to the density of the navigation subtree rooted at this node in terms of citation count. Intuitively, the selection is done such that every expansion reduces maximally the expected remaining navigation cost. For example, to reach the concepts “Cell Proliferation” and “Apoptosis” using the BioNav navigation method 19 concepts are revealed, after 5 node expansions, compared to 123 concepts, also after 5 expansions, with the static navigation method of Fig. 1.

In addition to the static hierarchy navigation works mentioned above, there are works (e.g., the Clusty search engine [22], or [2], [3]) on dynamic categorization of query results, which create unsupervised query-dependent results clusters. 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.

We make the following contributions:

1. A comprehensive framework for navigating large query results using extensive concept hierarchies (Section II).
2. A formal cost model for measuring the navigation cost paid by the user (Sections III and IV).
3. A complexity result proving that expanding the tree in a way that minimizes the user’s navigation cost is an NP-complete problem (Section V).

4. An efficient heuristic and a feasible optimal algorithm for minimizing the navigation cost (Section VI).
5. Experimental results validating the effectiveness of the BioNav system when compared to state-of-the-art categorization systems (Section VIII).
6. An online version of the BioNav system is available at <http://db.cse.buffalo.edu/bionav>.

The BioNav system architecture and implementation is presented in Section VII. Related work is discussed in Section IX and the paper concludes in Section X.

II. FRAMEWORK AND BIONAV OVERVIEW

The MeSH concept hierarchy is the starting point of the framework and is defined as follows.

Definition 1 (Concept Hierarchy): A *Concept Hierarchy* $H(V, E, r)$ is a labeled tree consisting of a set V of concept nodes, a set E of edges and is rooted at node r . Each node $n \in V$ has a label l and a unique identifier id . \square

According to the semantics of the MeSH concept hierarchy [13], the label of a child concept node is more specific than the one of its parent. This also holds for most concept hierarchies.

Once the user issues a keyword query, PubMed–BioNav uses the Entrez Programming Utilities (eUtils) [6]–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. Formally, an Initial Navigation Tree $T_i(V_i, E_i, r)$ is a concept hierarchy, where every node (concept) $n \in V_i$ is additionally labelled with a *results (citations) list* $L(n)$.

In a given initial navigation tree, several concept nodes might have an empty results list. Since MeSH is a rather large concept hierarchy, BioNav reduces the size of the initial navigation tree by removing the nodes with empty results lists, while preserving the ancestor/descendant relationships. Formally, the resulting structure is defined as follows.

Definition 2 (Navigation Tree): A *Navigation Tree* $T(V, E, r)$ is the maximum embedding of an initial navigation tree $T_i(V_i, E_i, r)$ such that no node $n \in V$ is labeled with an empty results list $L(n)$, excluding the root (in order to maintain the tree structure and avoid the creation of a forest). \square

In principal, an *embedding* $T(V, E, r)$ of a tree $T_i(V_i, E_i, r)$ is an injection from V to V_i such that every edge in E corresponds to a path (disjoint from all other such paths) in T_i . An embedding T of a tree T_i , where both trees are rooted at node r , is *maximum* if no other node n with a nonempty results list $L(n)$ can be added to V and T still be an embedding. The maximum embedding of the initial navigation tree is recursively computed in a single depth-first left-to-right traversal. If a node n has an empty results list $L(n)$, then replace n with its children. If n is a leaf, then simply remove it. Fig. 3 shows part of the navigation tree for the “prothymosin” query, where the results lists are omitted for clarity.

The above procedure reduces the size of the initial navigation tree, but the structure is still too big (3,940 nodes for query “prothymosin”) to simply display it to the user or let

her navigate it, especially if her query is of exploratory nature. BioNav minimizes the user’s effort to reach the desired citations in the navigation tree by expanding in a way that minimizes the expected overall user navigation cost. Moreover, BioNav avoids information clutter by hiding unimportant concept nodes leading to interesting ones. This is achieved through a series of *expand* actions that reveal only a few descendants (not necessarily children) of the user selected node for further navigation.

We model a node expansion at a given navigation step as an *EdgeCut* in the navigation tree. In graph theory, an EdgeCut in a graph $G(V, E)$ is a set of edges $E_C \subseteq E$ such that the graph $G'(V, E \setminus E_C)$ has more components than G . For trees, any subset of the edges constitutes an EdgeCut, since the removal of any single edge creates a new component.

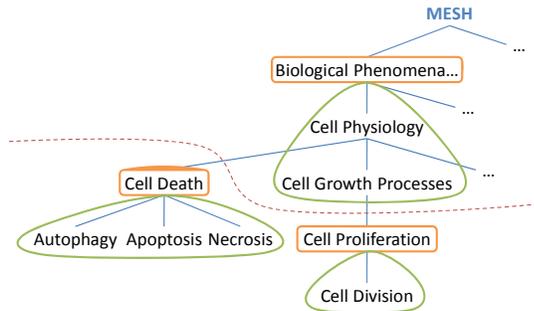


Fig. 3 Navigation Tree, EdgeCut and Component Subtrees

In Fig. 3, the dashed line illustrates the EdgeCut corresponding to the expansion of the node “Biological Phenomena...”. This expansion reveals the highlighted concepts of Fig. 3, which include a subset of the highlighted concepts in Fig. 2c. The EdgeCut consists of the edges (“Cell Physiology”, “Cell Death”) and (“Cell Growth Processes”, “Cell Proliferation”). Intuitively, an EdgeCut allows us to “skip” child nodes and navigate directly to descendent nodes located deeper in the tree and show them as children of the node being expanded. Moreover, an EdgeCut can selectively reveal only a subset of a potentially large set of descendent nodes, as is the case in Fig. 2c where only 5 out of the 25 descendants of “Biological Phenomena...” are revealed.

Definition 3 (Valid EdgeCut): A valid EdgeCut of a tree $T(V, E, r)$ is an EdgeCut $C \subseteq E$ such that no two edges in C appear in a path from the root to a leaf node. □

We only consider valid EdgeCuts in the rest of the paper, because invalid EdgeCuts lead to unintuitive navigations.

Component Subtrees An EdgeCut causes the creation of two types of *component subtrees*, *upper* and *lower*. Given an EdgeCut C of a tree $T(V, E, r)$, a *lower* component subtree rooted at y_i is created by each node $y_i \in V$, such that $(x, y_i) \in C$ for some node x . In Fig. 2c, at the expansion of node “Biological Phenomena...”, five lower component subtrees are created, two of which are shown in Fig. 3, rooted at “Cell Death” and “Cell Proliferation”. Moreover, for a given EdgeCut C , a single *upper* component subtree is created and comprises of the nodes not in any *lower* component subtree, and is always rooted at the root of the tree being expanded. In Fig. 3, the upper component subtree comprises of the nodes

“Biological Phenomena...” (root), “Cell Physiology” and “Cell Growth Processes”.

The state of the navigation tree after an EdgeCut, and the component subtrees created, is captured by the *Active Tree* defined below.

Definition 4 (Active Tree): An *Active Tree* $T_A(V, E, r)$ is a Navigation Tree where each node $n \in V$ is annotated with a node set $I(n)$ consisting of the nodes in the component subtree rooted at n . If a node n is not a root of a component subtree, then $I(n) = \{n\}$. The non-singleton I sets are disjoint. □

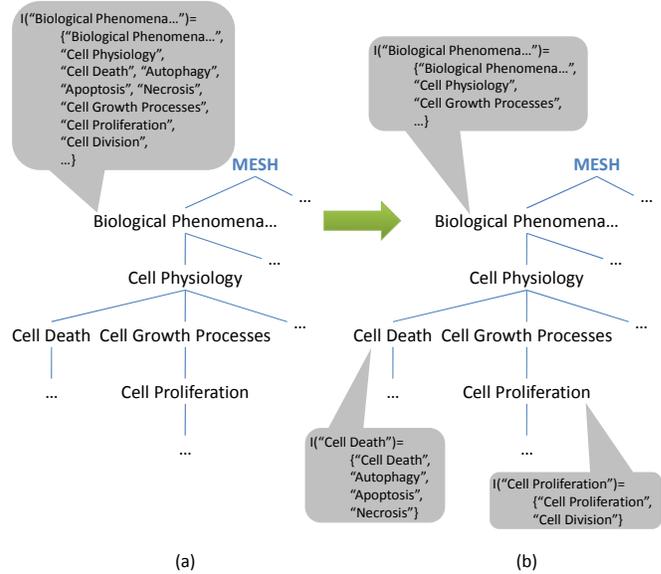


Fig. 4 The Active Tree before and after the EdgeCut operation in Fig. 3

Before any EdgeCut, a navigation tree is trivially converted to an active tree by annotating the root node with an I set that includes all tree nodes. The rest of the nodes n_i are annotated with the node set $I(n_i) = \{n_i\}$. Fig. 4a shows (part of) the active tree capturing the state of the navigation tree before the EdgeCut in Fig. 3 (singleton I sets are not shown).

An EdgeCut (expansion) is an operation on the active tree, performed on the I set of a given node, and updates the sets $I(n_i)$ of the roots n_i of the upper and lower subtrees created by the EdgeCut based on the nodes included in these subtrees. The operation is denoted by $\text{EdgeCut}: I(n) \rightarrow S \subseteq I(n)$ and returns the set S of roots of the upper and lower subtrees that it creates. Fig. 4b shows the effect of the EdgeCut operation in Fig. 3 on the active tree in Fig. 4a. The active tree is closed under the EdgeCut operation.

Note that the set $I(n)$ of a node n is overloaded to also denote the “invisible” component subtree of the active tree that is rooted at n and only consists of the nodes in $I(n)$. For instance, the invisible subtree I (“Biological Phenomena...”) in Fig. 4b is the one indicated as the upper component subtree in Fig. 3.

BioNav visualizes the active tree to the user by showing only the nodes that do not appear in any non-singleton I set organized according to the following definition.

Definition 5 (Active Tree Visualization): The visualization of an active tree $T_A(V, E, r)$ is the embedded tree $T_A'(V', E', r)$, where V' consists of the nodes not in any non-singleton $I(n)$, for all $n \in V$. Shown next to every node $n \in V'$ is the number of distinct citations attached to nodes in $I(n)$, given by $|L(I(n))| = |\bigcup_{n_i \in I(n)} L(n_i)|$. If n has a non-singleton $I(n)$, then an *expand* hyperlink is shown next to it. \square

The visualization of the active tree after the EdgeCut in Fig. 3 is shown in Fig. 2c. Note that the citation count $|L(I(n))|$ for “Biological Phenomena...” in Fig. 2c is 166 denoting the unique citations attached to it and its (invisible) component subtree. It is reduced from 217 in Fig. 2b, since its component subtree is getting smaller as more concept nodes are revealed.

An EdgeCut and the visualization of the resulting active tree are capable of reducing the navigation tree both height- and width-wise. The embedded tree in Fig. 2c, compared to the navigation tree in Fig. 1, is narrower and shorter.

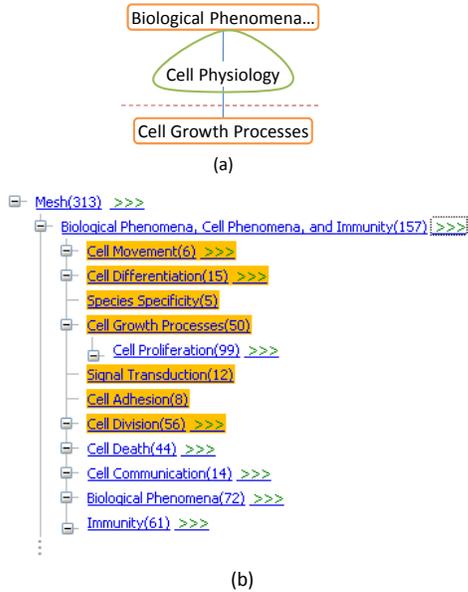


Fig. 5 (a) EdgeCut operation on the upper subtree “Biological Phenomena...” and (b) the visualization of the resulting active tree

Using the “>>>” hyperlinks, the user can trigger subsequent EdgeCut operations on component subtrees in a recursive fashion. Although we expect the user to trigger EdgeCut operations predominantly on the lower component subtrees, an EdgeCut is possible on the upper subtree as well. Fig. 5a shows an EdgeCut operation on the upper component subtree of Fig. 3 that reveals the “Cell Growth Processes” concept. Fig. 5b displays the visualization of the resulting active tree where “Cell Growth Processes” is now the parent of the previously revealed concept “Cell Proliferation”.

III. NAVIGATION AND COST MODEL

The navigation model of BioNav is formally defined in this section. Then the navigation cost model is presented, which is used to devise and evaluate our algorithms in later sections.

Navigation Model After the user issues a keyword query, BioNav initiates a 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 $I(n)$ rooted at concept node n :

1. **EXPAND $I(n)$:** The user clicks on the “>>>” hyperlink next to node n and causes an $\text{EdgeCut}(I(n))$ 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 EdgeCut operation.

This navigation process continues until the user finds *all* the citations she is interested in.

In order to define a cost model, we focus on a simplification of the general navigation model, which we call TOPDOWN, where only EXPAND, SHOWRESULTS and IGNORE are the available operations, that is, the user follows a top-down only navigation starting from the root. TOPDOWN is common in practice. Note that when the user encounters a leaf node in TOPDOWN the only available option is SHOWRESULTS. The TOPDOWN navigation model is formally presented in Fig. 6. It is a recursive procedure and is initially called on the root of the initial active tree.

```

EXPLORE( $I(n)$ )
if  $n$  is the root
   $S \leftarrow \text{EXPAND } I(n)$  // that is  $S \leftarrow \text{EdgeCut}(I(n))$ 
  For each  $n_i$  in  $S$ 
    EXPLORE( $I(n_i)$ )
else, if  $n$  is not a leaf-node, then choose one of the following:
1. SHOWRESULTS  $I(n)$ 
2. IGNORE  $I(n)$ 
3.  $S \leftarrow \text{EXPAND } I(n)$ 
   For each  $n_i$  in  $S$ 
     EXPLORE( $I(n_i)$ )
else, choose one of the following: //  $n$  is a leaf node
1. SHOWRESULTS  $I(n)$ 
2. IGNORE  $I(n)$ 

```

Fig. 6 TOPDOWN Navigation Model

TOPDOWN Cost Model The cost model, which is inspired by [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 SHOWRESULTS action. In particular, the cost model assigns (i) cost of 1 to each newly revealed concept node that the user examines after an EXPAND action, (ii) cost of 1 to each EXPAND action the user executes, and (iii) cost of 1 to each citation displayed after a SHOWRESULTS action.

For example, in the navigation of Fig. 2 above, the cost for reaching the “Cell Proliferation” concept and inspecting its

attached citations is 119. That is, 3 EXPAND actions on the root that reveal a total of 11 concept nodes, 1 EXPAND action on the “Biological Phenomena...” concept that reveals 5 nodes, and a SHOWRESULTS action on the “Cell Proliferation” concept that lists 99 citations. The user examines all concept nodes and all citations in order to select the ones of interest.

Since the exact sequence of actions of a user cannot be known *a priori*, we *estimate* the cost based on the following two probabilities:

- EXPLORE probability $P_E(I(n))$ is the probability that the user is interested in the component subtree $I(n)$ and will hence explore it. The IGNORE probability is $1 - P_E(I(n))$.
- EXPAND probability $P_C(I(n))$ is the probability that the user executes an EXPAND action on component subtree $I(n)$ given that she has chosen to explore $I(n)$. The SHOWRESULTS probability for $I(n)$ is $1 - P_C(I(n))$.

In Section IV, we show how we estimate probabilities $P_E(I(n))$ and $P_C(I(n))$. The cost of exploring component subtree $I(n)$, rooted at node n , is:

$$\text{cost}(I(n)) = P_E(I(n)) \cdot \left((1 - P_C(I(n))) \cdot |L(I(n))| + P_C(I(n)) \cdot \left(1 + |S| + \sum_{s \in S} \text{cost}(I_C(s)) \right) \right)$$

The first operand of the addition inside the big parenthesis is the cost of executing SHOWRESULTS on n . The second operand is the cost of executing an EXPAND action on n . The constant 1 is the cost of executing the EXPAND action, and S is the set of concept nodes revealed by the action, or otherwise the roots of component subtrees returned by the EdgeCut operation. $I_C(s)$ is the updated I set of a node $s \in S$ after the EXPAND action on $I(n)$ has been performed.

Recall that $|L(I(n))|$ in the cost formula is the number of distinct citations attached to $I(n)$. Intuitively, creating a component subtree with large number of duplicates reduces the navigation cost if the SHOWRESULTS probability for that subtree is high. Moreover, the number of duplicates across component subtrees should be minimal; otherwise the user will pay the cost of inspecting a citation multiple times.

Finally, note that by changing the cost assigned to executing an EXPAND action (which we set to 1 above) we affect the number of revealed concepts after each EXPAND. In particular, increasing this cost leads to more concepts revealed for each EXPAND.

IV. ESTIMATION OF NAVIGATION PROBABILITIES

We assume that each citation is *equally likely* to be of interest to the user. If more information about the “goodness” of the citations were available, our approach could be straightforwardly adapted using appropriate weighting for $L(I(n))$.

Estimating EXPLORE Probability P_E Since all citations in the query result are assumed to be of equal importance, concept n is of higher interest if $L(n)$ is large. On the other hand, a concept that is associated with a very large number of

citations $L_T(n)$ of MEDLINE, independently of the query, is probably not discriminatory or important. The latter is inspired by the inverse document frequency measure in Information Retrieval. Hence, $P_E(n)$ is proportional to $|L(n)|/|L_T(n)|$. We normalize $P_E(n)$ by dividing by the sum of all P_E 's in the active tree, that is:

$$P_E(n) = \frac{|L(n)|}{|L_T(n)|} \frac{1}{\sum_{n_i \in T} \frac{|L(n_i)|}{|L_T(n_i)|}}$$

For a component tree $I(n)$ rooted at node n :

$$P_E(I(n)) = \sum_{n_i \in I(n)} P_E(n_i)$$

Given the above formula, for the initial active tree it is $P_E(I(n)) = 1$. The above P_E formulas, together with the cost model in Section III, largely determine the characteristics of the component subtrees BioNav creates during an EXPAND action. In particular, the upper component subtree typically groups together (i) concepts with low P_E and a large number of attached citations, and (ii) concepts with high P_E and a small number of attached citations. The first group is dismissed as uninteresting and the second could lead to a large number of concepts being revealed. Intuitively, the two groups of concepts average each other out according to the $P_E(I(n))$ formula. The lower component subtrees typically group concepts with P_E and number of attached citations in-between the two extremes in a way that minimizes the average navigation cost.

Estimating EXPAND Probability P_C $P_C(I(n))$ is 0, if n is a leaf concept node or has a singleton $I(n)$ set, since there is no other choice for the user. For internal nodes in the active tree with a non-singleton $I(n)$ set that have a large $L(I(n))$, a typical user will want to further narrow down when faced with the prospect of seeing too many citations, that is, $P_C(I(n))$ is 1, if $L(I(n))$ is greater than an *upper threshold*. $P_C(I(n))$ is 0, if $L(I(n))$ is smaller than an *lower threshold*. Currently, BioNav operates with 50 and 10 being the upper and lower threshold respectively.

In the remaining cases, a user might want to narrow down the search of $I(n)$, by executing an EXPAND action, if the citations under n are widely distributed among the subconcepts in $I(n)$. An objective measure for such a wide distribution (disorder) is *information entropy*. If the entropy of the subtree $I(n)$ is large, then the user would benefit by an EXPAND action. So, $P_C(I(n))$ is computed as follows:

$$P_C(I(n)) = E(I(n)) = \frac{-\sum_{n_i \in I(n)} \frac{|L(n_i)|}{|L(I(n))|} \log \frac{|L(n_i)|}{|L(I(n))|}}{-\log \frac{1}{|I(n)|}}$$

The sum can become greater than 1 because of the existence of duplicates. Hence, we normalize the entropy of $I(n)$ by dividing with the maximum entropy, where citations are uniformly distributed to all nodes in $I(n)$ and there are no duplicates.

P_C determines the impact of duplicates in a component subtree after a node expansion. If $P_C(I(n))$ is low, that is, the SHOWRESULTS probability is high, then the number of

duplicates in $I(n)$ plays a bigger role in the way a component subtree is expanded.

V. COMPLEXITY RESULTS

To prove that the problem of selecting the optimal valid EdgeCut for a given tree is NP-Hard, where “optimal” means minimize the user navigation cost according to the navigation model of Section III, we prove that the problem is NP-complete for a simplified navigation model, which we refer to as TOPDOWN-EXHAUSTIVE and is a special case of the TOPDOWN model shown in Fig. 6.

In TOPDOWN-EXHAUSTIVE, BioNav performs an EXPAND action (EdgeCut operation) on the root of the initial active tree, and then the user selects randomly the root of one of the component subtrees created and performs a SHOWRESULTS action. The cost of TOPDOWN-EXHAUSTIVE navigation is the cost to read the root label of all component subtrees revealed by the EdgeCut plus the cost of SHOWRESULTS for the selected component subtree.

Intuition on the complexity of computing optimal valid EdgeCut: The “optimal” valid EdgeCut is the EdgeCut that will lead to the minimum expected navigation cost, that is, the minimum average cost. In order to minimize the expected cost of TOPDOWN-EXHAUSTIVE navigation, we need to minimize the cost of EXPAND and of SHOWRESULTS. The cost of EXPAND is simply the number k of component subtrees produced by the EdgeCut. The average cost of SHOWRESULTS over all component subtrees equals the sum of unique elements (citations) in every subtree over k . This sum would be $|L(T)|$ where T is the navigation tree if there were no duplicates among the subtrees. However, due to the existence of duplicates (the same citation can be annotated with multiple MeSH concepts) this sum depends on the EdgeCut. Hence, the duplicates are the reason that the problem is NP-complete for TOPDOWN-EXHAUSTIVE, because we need to maximize the number of duplicates within the created subtrees, and at the same time create a relatively small number of component subtrees. Note that even for a given k , the problem of selecting the best EdgeCut is NP-hard as we show in Theorem 1.

Theorem 1: Finding the optimal valid EdgeCut in TOPDOWN-EXHAUSTIVE is NP-complete. \square

Proof: The decision problem corresponding to the problem of computing the optimal EdgeCut is the following:

TOPDOWN-EXHAUSTIVE Decision (TED) Problem: Given a navigation tree T , where each node n contains a list $L(n)$ of elements from universe U (U are all the citations in the query result), that is, $L(n) \subseteq U$, there exists an EdgeCut C of T that creates k subtrees (including the upper subtree) with d duplicate elements within the created subtrees. That is, if S_1, \dots, S_k are the subtrees and each S_i contains $b(S_i)$ duplicates (if an element appears 3 times, then it counts as 2 duplicates), then $\sum_{i=1..k} b(S_i) = d$.

Note that the cost of a TOPDOWN-EXHAUSTIVE navigation is computed as follows, if we solve the TED problem for every combination of k and d . If T has W unique results, then a subtree of the EdgeCut will have on average

$(W + d)/k$ results. Hence the whole navigation cost is $k + (W + d)/k$, where k is the cost of reading the labels of the k subtrees.

TED is in NP since a solution can be verified in polynomial time. To prove that it is NP-complete, we will reduce the MAXIMUM EDGE SUBGRAPH (MES) problem, which is NP-complete [7], to TED.

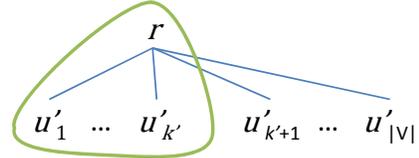
MAXIMUM EDGE SUBGRAPH (MES) Problem: Given graph $G(V, E)$, a weight function $w: E \rightarrow N$ (N are the natural numbers) and a positive integer k' , there is a subset $V' \subseteq V$ with $|V'| = k'$ such that the sum of the edge weights of the edges between the nodes in V' is d , that is, $\sum_{(u,v) \in E \cap (V' \times V')} w(u, v) = d$.

Mapping of MES to TED: For each node $u \in V$, we create a node u' in T that is a child of the root of T . That is, the root r of T is empty $L(r) = \emptyset$ and it has $|V|$ children.

The universe U is defined as follows: for each pair of edges $(u, v) \in E$ with weight $w(u, v)$, we add elements $B_{uv}^1, \dots, B_{uv}^{w(u,v)}$ in U .

Each of the nodes of T is populated with elements from U as follows: For each edge $(u, v) \in E$, we add to nodes u' and v' of T the elements $B_{uv}^1, \dots, B_{uv}^{w(u,v)}$. The intuition is that we map an edge weight in MES to number of duplicates between two nodes in TED.

We set $k = |V| - k' + 1$. In the figure below, the EdgeCut splits the tree into k subtrees.



Now, a solution to MES is mapped to a solution to TED, since selecting k' nodes in MES corresponds to expanding the tree into k subtrees in TED. The nodes of V corresponding to the nodes in the upper subtree of the EdgeCut (the one including the root) are the solution to MES. This set of nodes has maximum sum of edge weights in MES and maximum number of duplicates in TED.

Note: We assume that a node in TED can have the same element L multiple times. We could raise this assumption and just replace this node with a subtree that contains the element L multiple times. We did not do so to simplify the presentation of the proof. \square

VI. ALGORITHMS FOR BEST EDGE CUT

Given the cost equation in Section III, we can compute the optimal cost by recursively enumerating all possible sequences of valid EdgeCuts, starting from the root and reaching every concept in the navigation tree, computing the cost for each step and taking the minimum. However, this algorithm is also prohibitively expensive. Instead we propose an alternative algorithm *Opt-EdgeCut* that makes use of the *dynamic programming* technique to reduce the computation cost. As shown in Section VI-A below, *Opt-EdgeCut* is still

exponential and is just used to evaluate the quality of the heuristic we present in Section VI-B.

A. Optimal Algorithm for Best EdgeCut

The *Opt-EdgeCut* algorithm to compute the minimum expected navigation cost (and the EdgeCut that achieves it) traverses the navigation tree in post-order and computes the navigation cost bottom-up starting from the leaves. For each node n , the algorithm enumerates and stores the list $\mathbb{C}(n)$ of all possible EdgeCuts for the subtree rooted at n , and the list $\mathbb{I}(n)$ of all possible $I(n)$ sets that node n can be annotated with. The inclusion-exclusion principle [4] is used when enumerating $\mathbb{C}(n)$ and $\mathbb{I}(n)$, which leads to an ordering that maximizes reuse in the dynamic programming algorithm. The algorithm then computes the minimum cost for each subtree in $\mathbb{I}(n)$ given the EdgeCuts in $\mathbb{C}(n)$ and the already computed minimum costs for the descendants of n . The complexity of *Opt-EdgeCut* is $O(|V| \cdot 2^{|E|})$.

Algorithm *Opt-EdgeCut*

Input: The navigation tree T

Output: The best EdgeCut

```

1 Traversing  $T$  in post-order, let  $n$  be the current node
2 while  $n \neq \text{root}$  do
3   if  $n$  is a leaf node then
4      $\text{mincost}(n, \emptyset) \leftarrow P_E(n) * L(n)$ 
5      $\text{optcut}(n, \emptyset) \leftarrow \{\emptyset\}$ 
6   else
7      $\mathbb{C}(n) \leftarrow$  enumerate all possible EdgeCuts
       for the tree rooted at  $n$ 
8      $\mathbb{I}(n) \leftarrow$  enumerate all possible subtrees
       for the tree rooted at  $n$ 
9     foreach  $I(n) \in \mathbb{I}(n)$  do
10      compute  $P_E(I(n))$  and  $P_C(I(n))$ 
11      foreach  $C \in \mathbb{C}(n)$  do
12        if  $C$  is a valid EdgeCut for  $I(n)$  then
13           $\text{cost}(I(n), C) \leftarrow$ 

$$P_E(I(n)) \cdot \left( \begin{array}{l} (1 - P_C(I(n))) \cdot L(I(n)) \\ + P_C(I(n)) \cdot (1 + |S| + \sum_{s \in S} \text{mincost}(I_C(s))) \end{array} \right)$$

14        else
15           $\text{cost}(I(n), C) = \infty$ 
16         $\text{mincost}(n, I(n)) \leftarrow \min_{C_i \in \mathbb{C}(n)} \text{cost}(I(n), C_i)$ 
17         $\text{optcut}(n, I(n)) \leftarrow C_i$ 
18 return  $\text{optcut}(\text{root}, E)$  //  $E$  is the set of all tree edges
```

B. Heuristic-ReducedOpt Algorithm

The algorithm to compute the optimal navigation, *Opt-EdgeCut*, is exponential and hence infeasible for the navigation trees of most queries. We propose a heuristic to select the best EdgeCut for a node expansion. Note that the input argument to the heuristic is a component tree $I(n)$ and not the whole active tree T as in *Opt-EdgeCut*. The reason is that once *Opt-EdgeCut* is executed for T , the costs (and optimal EdgeCuts) for all possible $I(n)$'s are also computed and hence there is no need to call the algorithm again for subsequent expansions.

For a given component subtree $I(n)$, *Opt-EdgeCut* enumerates a large number of EdgeCuts on $I(n)$ and repeats this recursively on its subtrees.

We propose to run *Opt-EdgeCut* on a reduced version $I'(n)$ of $I(n)$. The reduced tree $I'(n)$ has to be small enough so that *Opt-EdgeCut* can run on it in "real-time". Also, $I'(n)$ should approximate $I(n)$ as closely as possible. $I'(n)$ is the tree of "supernodes" created by partitioning $I(n)$. Each supernode in $I'(n)$ corresponds to a partition of tree $I(n)$. Then, *Opt-EdgeCut* is executed on $I'(n)$.

The algorithm we use to partition the tree is based on the k -partition algorithm [11] that processes the tree in a bottom-up fashion. For each tree node n , the algorithm removes the "heaviest" children of n one-by-one until the weight of n falls below k . For each of the removed children, it creates a partition. The result is a tree-partitioning with the minimum cardinality. The complexity of the k -partition algorithm is $O(|V| \cdot \log|V|)$.

We adopt the k -partition algorithm to our needs as follows. For each node in $I(n)$, we assign weight equal to $|L(n)|$. We run the k -partition algorithm by setting k , the weight threshold, to $\sum_{n_i \in I(n)} L(n_i)/z$, where z is the number of desired partitions. However, this might result in more than z partitions, due to some non-full partitions. Therefore we repeatedly run k -partition algorithm on $I(n)$, gradually increasing k until up to z partitions are obtained. Note that z is the maximum tree size on which *Opt-EdgeCut* can operate in "real-time". In our experiments we set $z = 10$.

Algorithm *Heuristic-ReducedOpt*

Input: Component subtree $I(n)$, number z of partitions

Output: The best EdgeCut

```

1.  $z' \leftarrow z$ 
2. repeat
3.    $k \leftarrow \sum_{n \in T} L(n)/z'$ 
4.    $\text{Partitions} \leftarrow k\text{-partition}(I(n), k)$ 
   // call  $k$ -partition algorithm [11]
5.    $z' \leftarrow z' - 1$ 
6. until  $|\text{Partitions}| \leq z$ 
7. construct reduced subtree  $I'(n)$  from  $\text{Partitions}$ 
8.  $\text{EdgeCut}' \leftarrow \text{Opt-EdgeCut}(I'(n))$ 
9.  $\text{EdgeCut} \leftarrow$  corresponding of  $\text{EdgeCut}'$  for  $I(n)$ 
10. return  $\text{EdgeCut}$ 
```

VII. SYSTEM ARCHITECTURE AND IMPLEMENTATION

The BioNav system architecture is shown in Fig. 7 and consists of two parts. The off-line components populate the BioNav database with the MeSH concept hierarchy and the associations of the MEDLINE citations with MeSH concepts, while the on-line components support BioNav's web interface and the EXPAND/SHOWRESULTS actions of the user.

Off-Line Pre-Processing The BioNav database is first populated with the MeSH hierarchy, which is available online [13]. There are more than 48,000 concept nodes in the hierarchy.

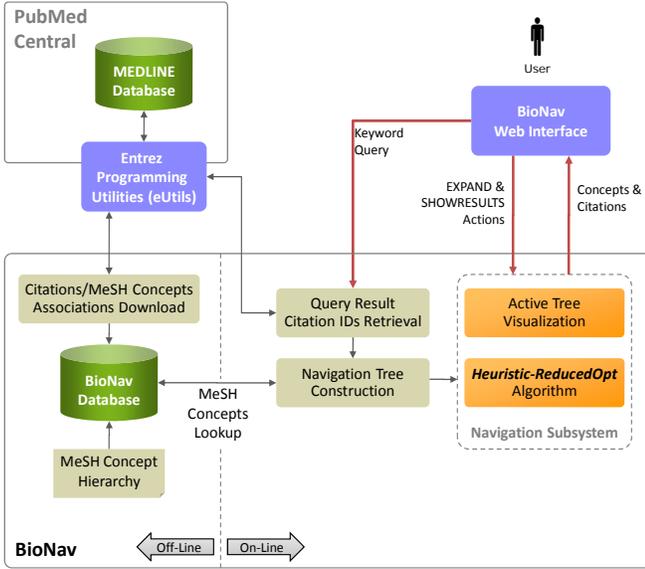


Fig. 7 BioNav System Architecture

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 a tuple $\langle \text{concept}, \text{citationID} \rangle$ to a table in the BioNav database. Alternatively, we could determine the associations by using the MeSH concepts that each citation is annotated with in 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 $\langle \text{concept}, \text{citationID} \rangle$ tuples. In the end, there were almost 747 million such tuples. To improve the selection queries on this table, we de-normalized it by concatenating all concepts associated with each citation into a comma-separated list, that is:

$$\langle \text{citationID}, (\text{concept1}, \text{concept2}, \dots) \rangle$$

Hence, the size of the associations table became as big as the number of citations in MEDLINE.

When executing the queries using the concepts as keywords, we also store the number of citations $L_T(n)$ in the query result, since it is needed for the computation of P_E in Section IV.

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 (PubMed Identifiers) of the citations in the query result. This is done

using the ESearch utility of the Entrez Programming Utilities (eUtils) [6]. 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.

The navigation tree is trivially converted to an active tree (see Section II) and passed on the *Navigation Subsystem* that supports the user's actions on the BioNav web interface. Initially, the navigation subsystem just visualizes the active tree on the web interface, that is, it simply shows its root node. Subsequently, the user requests an EXPAND action on the root. Then, the navigation subsystem executes the *Heuristic-ReducedOpt* algorithm on the tree $T(r)$ of the root r , and the resulting active tree is visualized on the web interface.

When the user makes a SHOWRESULTS request, BioNav uses the Entrez ESummary utility to download high level information of the citations to be shown, such title and authors.

VIII. EXPERIMENTAL EVALUATION

We evaluated the BioNav system in terms of both average navigation cost and expansion time performance. In particular, we first (Section VIII-A) show that the BioNav navigation method, which is evaluated using the *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. Note that the optimal algorithm, *Opt-EdgeCut*, was not evaluated, because its execution times are prohibiting even for relatively small (e.g., 30 nodes) navigation trees. Then (Section VIII-B) we show that the execution time of *Heuristic-ReducedOpt* is small enough to facilitate interactive-time user navigation.

We used the MEDLINE database, and BioNav was implemented as explained in Section VII. The experiments were executed on a Dell Optiplex machine with 3.2Ghz CPU and 2 GB of main memory, running Windows XP Professional. All algorithms were implemented in Java and compiled using Sun's JRE-1.6.2. We used the Oracle 10i database system to store the BioNav database.

A. Evaluation of Reduction in Navigation Cost

To evaluate the navigation cost benefit of BioNav, we created, with the help of our collaborators from the biomedical domain, a set of queries covering a number of research fields that satisfy biomedical scientists within different disciplines but also queries that interest "non-expert" users. This query workload is show in Table I.

"Follistatin" and "LbetaT2" are terms that mainly interest scientists studying reproductive endocrinology and gynecology. The "dyslexia genetics" query accumulates results related to genes associated with dyslexia and interest geneticists but also psychologists. "Melibiose permease" and "Na⁺/I⁻ symporter" are transport proteins related to bacterial

TABLE I
QUERY WORKLOAD

Keyword(s)	# of Citations in Query Result	Navigation Tree Size	Maximum Tree Width	Maximum Tree Height	Tree Citations w/ Duplicates	Target Concept	MeSH Level of Target Concept	$ L(n) $ of Target Concept	$ LT(n) $ of Target Concept
LbetaT2	116	1947	1009	10	14927	Mice, Transgenic	5	11	90804
melibiose permease	160	1324	722	8	14419	Substrate Specificity	3	31	79470
varenicline	162	1830	962	6	11370	Nicotinic Agonists	7	81	18277
Na+/I symporter	163	2596	1367	6	17146	Perchloric Acid	3	7	4250
prothymosin	313	3941	2113	10	30897	Histones	4	15	22741
ice nucleation	474	3181	1776	9	27440	Plants, Genetically Modified	3	2	12330
vardenafil	486	3424	2014	8	40987	Phosphodiesterase Inhibitors	5	401	69984
dyslexia genetics	517	3056	1691	9	45079	Polymorphism, Single Nucleotide	4	18	18843
syntaxin 1A	1115	6589	3764	10	105503	GABA Plasma Membrane Transport Protein	7	11	650
follistatin	1183	6446	3656	10	102946	Follicle Stimulating Hormone	6	157	34540

growth and thyroid function respectively. On the other hand, “vardenafil” (Levitra), used for the treatment of erectile dysfunction, and “varenicline” (Chantix), used for quitting smoking, are two relatively new drugs that interest many users with basic biochemical or medical knowledge.

Moreover, we chose queries that either correlate with quite a few fields of research or concentrate in more specific topics. For example, the literature for “prothymosin”, although not particularly broad in number of citations in the query result (313), is associated with several topics such as cancer, cell proliferation, apoptosis, chromatin remodeling, transcriptional regulation and immunity. In contrast, “vardenafil” retrieves a higher number of citations (486) but the literature is mostly targeted to erectile dysfunction and hypertension. This fact is reflected on the navigation tree characteristics for the two queries, also shown in Table I. The navigation tree for “prothymosin” is bigger than the one for “vardenafil” in every respect, that is, tree size, maximum width and height.

For each query in Table I, a desirable MeSH concept (referred as “target concept”) was chosen as the target of a possible user navigation. The target concepts are among the ones involved in the research fields closely related to the keyword query and were chosen based on specific research interests. Their characteristics in Table I include their location depth in the MeSH hierarchy, the number of citations $|L(n)|$ attached to them, and their total number of citations $|L_T(n)|$ in MEDLINE.

In this experiment we assume that the user follows a top-down navigation where she always chooses the right node to expand in order to finally reveal the target concept. We compare the navigation cost of BioNav, where EXPAND is implemented using the *Heuristic-ReducedOpt* algorithm (with $z = 10$), to a static navigation algorithm, where EXPAND displays all children of the expanded concept (similarly to current systems like GoPubMed [21] and Amazon²).

Fig. 8 compares the navigation cost for these two methods. We observe that BioNav often improves the navigation cost by an order of magnitude. The average improvement of BioNav, over the static navigation, is 85%. The improvement is high regardless of the navigation tree characteristics (84% for “prothymosin”, 94% for “vardenafil”), and regardless of

the number of citations in the query result (80% for “LbetaT2”, 89% for “follistatin”). The smallest improvement was observed for “ice nucleation” (67%). The reason is that its target concept (Plants, Genetically Modified) has an extremely low $|L(n)| = 2$. Hence, its P_E is quite low and so it takes several EXPAND actions until BioNav reveals it.

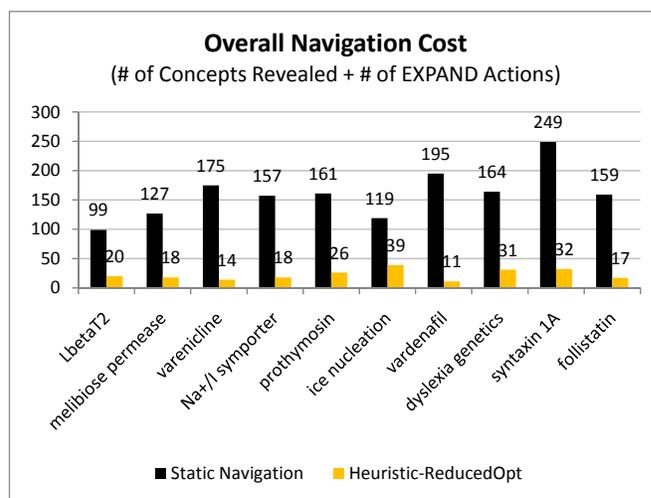


Fig. 8 Navigation Cost of BioNav vs. Static Navigation

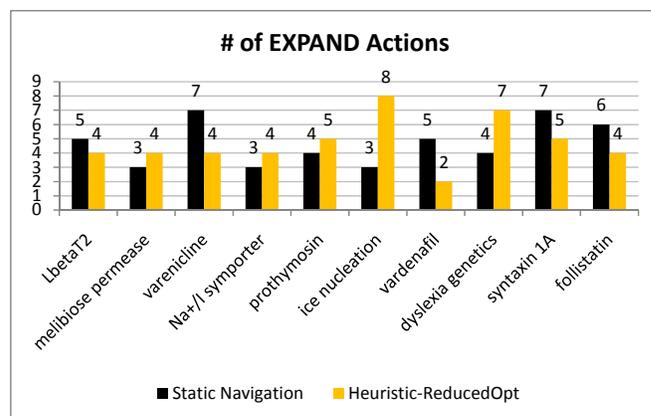


Fig. 9 EXPAND Action of BioNav vs. Static Navigation

² Even if we show a few children at a time and display a “more” button, the navigation cost does not considerably change, given that executing “more” incurs additional cost.

Fig. 9 shows the number of EXPAND actions for the two methods. Note that these numbers are relatively close for the two algorithms, which means that the dramatic differences in Fig. 8 are due to the fact that BioNav selectively reveals few descendant nodes for each EXPAND, instead of a possibly large number of child nodes. The worst case is the “ice nucleation” query, where BioNav requires 8 EXPAND actions, compared to 3 of static navigation, since the target concept is quite high in the MeSH hierarchy, and at the same time has a low P_E , as discussed above.

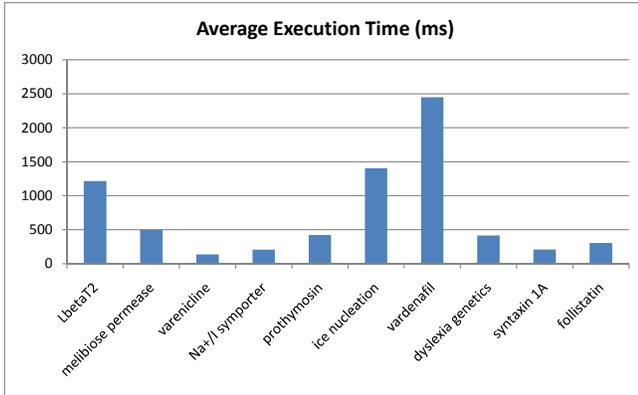


Fig. 10 *Heuristic-ReducedOpt* Performance

B. Evaluation of Execution Times for *Heuristic-ReducedOpt*

Fig. 10 shows the average time of *Heuristic-ReducedOpt* to execute an EXPAND action for each query in Table I. The average was taken over the number of EXPAND actions shown in Fig. 9. For an input tree $I(n)$, *Heuristic-ReducedOpt* first creates a reduced tree $I'(n)$, and then runs the *Opt-EdgeCut* algorithm on it. The execution time is dominated by *Opt-EdgeCut* as it is an exponential algorithm and depends on the size of the input tree. As stated earlier, we restrict the size of the reduced tree $I'(n)$ to 10 nodes. However, $I'(n)$ can have a smaller size (see Section VI-B), in which case *Opt-EdgeCut* runs much faster but with some loss in accuracy.

For example, the reduced tree $I'(n)$ for “vardenafil”, in both EXPAND actions, had size 10, which explains the highest average execution time, and also the highest improvement in Fig. 8. On the other hand, for “Na+/I symporter”, the first three EXPAND actions resulted in an $I'(n)$ of size 8, 7 and 9, respectively. Hence, the average execution time in Fig. 10 is lower, as is the improvement in navigation cost.

The size of $I'(n)$ is high if the distribution of citations in $I(n)$ is skewed, since fewer tree nodes can be placed within the same partition. Even if the size of $I'(n)$ is high though, the execution time can be low if $I'(n)$ is narrow. As an example, Fig. 11 shows the execution time of *Heuristic-ReducedOpt* for all 5 EXPAND actions of query “prothymosin”. Note that the fourth EXPAND results in an $I'(n)$ with 10 nodes and has a lower execution time than the first EXPAND, where the $I'(n)$ has 8 nodes. That’s because the MeSH hierarchy is wider on the upper levels. As the user navigates the tree top-down and reaches lower levels, $I'(n)$ becomes narrower.

An exception to the above behavior is the query for “ice nucleation”. The user keeps expanding the upper subtree to reach the high-level concept “Plants, Genetically Modified”, in which case $I'(n)$ is wide and its size close to 10 for all 8 expansions.

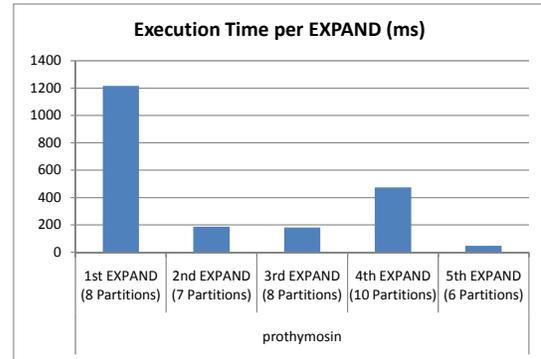


Fig. 11 *Heuristic-ReducedOpt* Performance for “prothymosin”

IX. RELATED WORK

Biomedical Search Systems Several systems have been developed to facilitate keyword search on PubMed using the MeSH concept hierarchy. The closest to BioNav is GoPubMed [5], [21], 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. We could not directly compare BioNav with GoPubMed in our experiments, since GoPubMed indexes citations differently than PubMed. However, the static navigation method we implemented and compared very closely approximates the behaviour and the navigation cost of using GoPubMed.

Other systems that tackle PubMed search using the MeSH concept hierarchy include PubMed PubReMiner [18] and XplorMed [16],[23]. Both of them are query refinement tools and do not implement a particular navigation method. In particular, PubMed PubReMiner outputs a long list of all MeSH concepts associated with each query along with their citation count. The user can select one or more of them and refine her query. XplorMed performs statistical analysis of the words in the abstracts of the citations in the query result and proposes query refinements/extensions to the user in a multi-step process. Ali Baba [9],[17] displays the results on a graph where edges denote associations between the result nodes, which are typically genes and proteins. iHOP [10] shows to the user the genes associated to a query gene, where the association is measured through co-occurrence in a sentence.

Hierarchical Results Navigation Systems In addition to GoPubMed discussed above, a few other systems offer hierarchical navigation on the query results. Amazon and eBay are the most popular systems that use static hierarchies to organize query results. Their static navigation method works relatively well since their hierarchies are significantly smaller than MeSH. BioNav could be applied on these hierarchies to minimize the expected navigation cost.

Two academic proposals [2]-[3] dynamically categorize SQL query results by inferring a hierarchy based on the characteristics of the result tuples. Their domain is the tuple attributes and their problem is how to organize them hierarchically in order to minimize the navigation cost. They also decide the value ranges for each attribute, for both categorical and numerical ones, and how to rank them. One of the systems [3] takes into consideration the user's preferences during the inference for a more personalized experience. Once the hierarchy is inferred, they follow a static navigation method. BioNav is distinct since it offers dynamic navigation on a predefined hierarchy, as is the MeSH concept hierarchy. Hence, BioNav is complementary to these systems, since it can be used to optimize the navigation, after these systems construct the navigation tree.

Clustering Systems Clustering systems [20], [22], [24] create unsupervised query-dependent results clusters. The Clusty [22] search engine clusters keyword-based query results on the web and operates on top of other search engines. HighWire Press [20] uses Clusty's algorithms to cluster query results in the biomedical domain. Once the clusters are created, a static navigation method is also followed. BioNav uses a cost model to minimize the navigation cost, while clustering systems minimize the distance between clusters. Again, BioNav could be adapted to work on top of the (typically shallow) hierarchy created by clustering systems.

X. CONCLUSIONS AND FUTURE WORK

Information overload is a major problem when searching biomedical databases like 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 descendent concepts, instead of simply showing all its children, ranked based on their estimated relevance to the user's query. We formally stated the underlying framework and the navigation and cost models used for the evaluation of our approach. 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. The architecture of the BioNav system was implemented and an online version is available at <http://db.cse.buffalo.edu/bionav>.

In the future, we plan to apply and adapt BioNav to other search systems that display a hierarchical view of the query results. We will look into various domains, including e-commerce and web search.

REFERENCES

- [1] 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 SQL-Query-Result Navigation*. SIGMOD Conference 2007: 641-652
- [4] Comtet, L.: *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] (2008) Entrez Programming Utilities. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
- [7] U. Feige, D. Peleg and G. Kortsarz: *The Dense k-Subgraph Problem*. Algorithmica 29 (2001) 410-421
- [8] V. Hristidis and Y. Papakonstantinou: *DISCOVER: Keyword Search in Relational Databases*. In Proc. of VLDB Conference, 2002
- [9] (2008) Humboldt-Universität zu Berlin – Ali Baba: PubMed as a graph. [Online]. Available: <http://alibaba.informatik.hu-berlin.de/>
- [10] (2008) iHOP - Information Hyperlinked over Proteins. [Online]. Available: <http://www.ihop-net.org/UniPub/iHOP/>
- [11] S. Kundu and J. Misra: *A Linear Tree Partitioning Algorithm*. SIAM J. Comput. 6(1): 151-154 (1977)
- [12] 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
- [13] Medical Subject Headings (MeSH®). <http://www.nlm.nih.gov/mesh/>
- [14] 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
- [15] (2008) OMIM - Online Mendelian Inheritance in Man. [Online]. Available: <http://www.ncbi.nlm.nih.gov/Omim/>
- [16] C. Perez-Iratxeta, P. Bork and M. A. Andrade: *Exploring MEDLINE Abstracts with XplorMed*. Drugs of Today. 2002;38:381-389
- [17] C. Flake, T. Schiemann, M. Pankalla, J. Hakenberg and U. Leser: *Ali Baba: PubMed as a graph*. Bioinformatics, 22(19):2444-2445, 2006
- [18] (2008) PubMed PubReMiner: A Tool for PubMed Query Building and Literature Mining. [Online]. Available: <http://bioinfo.amc.uva.nl/human-genetics/pubreminer/>
- [19] H. Shatkay, R. Feldman: *Mining the Biomedical Literature in the Genomic Era: An Overview*. Comput. Biol. 2003;10(6):821-55
- [20] (2008) Stanford University – HighWire Press. [Online]. Available: <http://highwire.stanford.edu/>
- [21] (2008) Transinsight GmbH – GoPubMed. [Online]. Available: <http://www.gopubmed.org/>
- [22] (2008) Vivísimo, Inc. – Clusty. [Online]. Available: <http://clusty.com/>
- [23] (2008) XplorMed: eXploring Medline abstracts. [Online]. Available: <http://www.ogic.ca/projects/xplormed/>
- [24] T. Zhang, R. Ramakrishnan and M. Livny: *BIRCH: An Efficient Data Clustering Method for Very Large Databases*. SIGMOD Conference 1996: 103-114