Optimizing Query Results using Middle Layers Based on Concept of Hierarchies

[1]K.S.N.V. Jyotsna Devi, [2]K. Rajini Kumari
[1,2]Dept. of CSE, A.S.R College of Engineering & Tech., Tetali, Tanuku, AP, 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 also be combined, have been proposed to alleviate this information overload problem. Result optimization and results categorization for biomedical databases is the focus of this work. 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

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 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. As an example [1] .
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 (shown in fig. 2) 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 II, describes related work, Section III, describes the Proposed System architecture Work nature, Section IV, presents experiment Evolution. Section V, presents conclusion and future scope

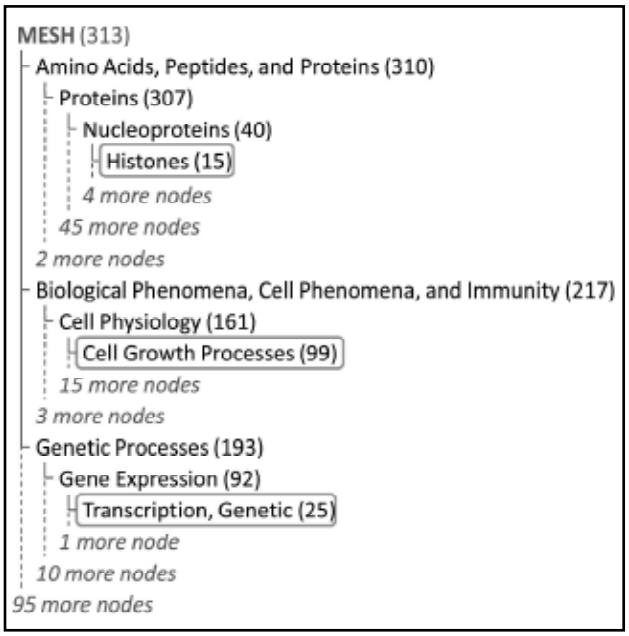

Fig. 1:

II. 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 [5], [28], 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 (e.g., MeSH [19]) 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 [1]—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 [1].

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 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.

Fig. 2a, for example, shows the initial expansion of the root node where only eight (highlighted) descendants are revealed compared to 98 children shown in Fig. 1. 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. 2b. 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. 2c, 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. 2c. In Fig. 2d, 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. 3. 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. 1. 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 . 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 sub tree 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, the reason that "Proteins" is not displayed in fig. 2, is that it is too general given the query results and the original distribution of citations in the PubMed database (details in Sections 3 and 4), and hence displaying it would lead to an expected increase in the user navigation cost, based on the user 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 [29], or [2-3]), 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. We make the following contributions:

1. A comprehensive framework for navigating large query results from PubMed using MeSH, an extensive concept hierarchy used for indexing citations in MEDLINE
2. A formal cost model for measuring the navigation cost incurred by the user.

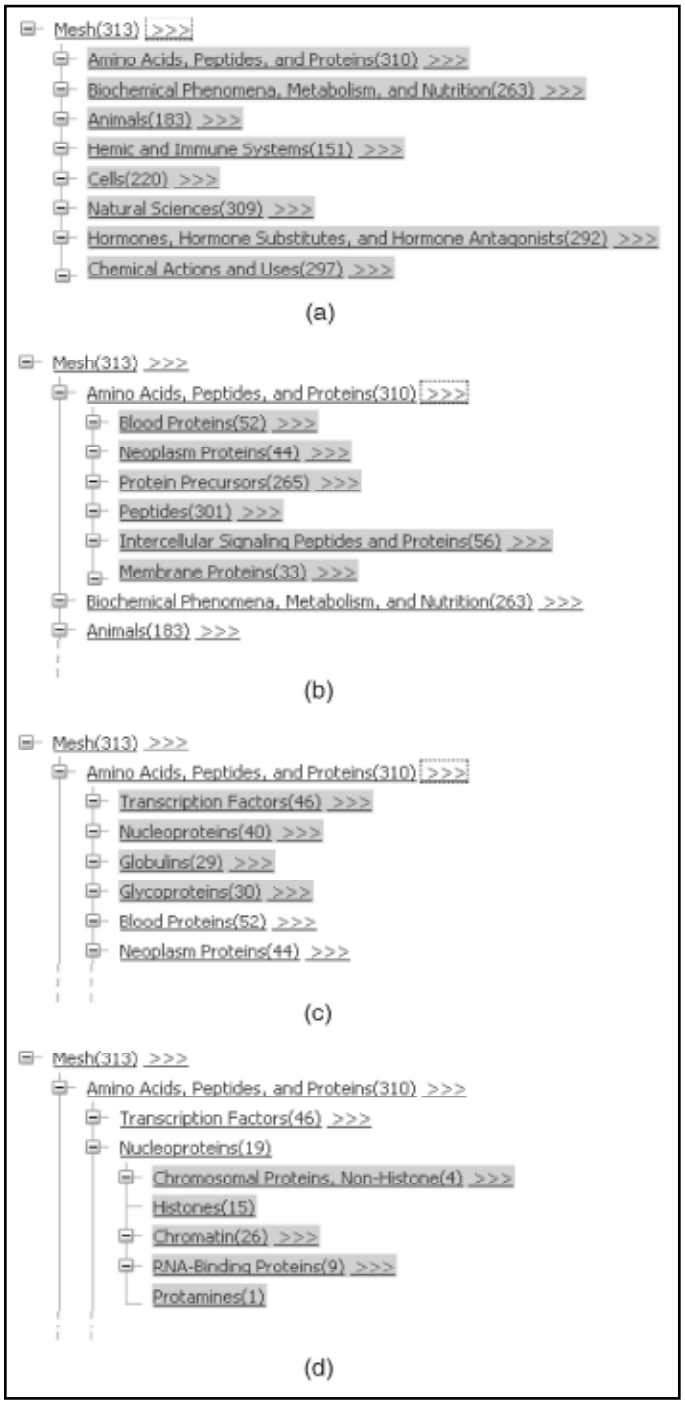

Fig. 2:

3. A complexity result proving that expanding the tree in a way that minimizes the user's navigation cost is an NP-complete problem.
4. An efficient heuristic and a feasible optimal algorithm for minimizing the navigation cost.
5. Experimental results validating the effectiveness of the BioNav system when compared to state-of-the art categorization systems.
6. An online version of the BioNav system is available at http://db.cse.buffalo.edu/bionav. Although we specifically target the biomedical domain in this work, the approach can be directly

applied to data sets where tuples are classified using terms from a concept hierarchy. The core of the first contribution has been presented in our preliminary short paper [13].

III. Framework and BioIntelR 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.

A. 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

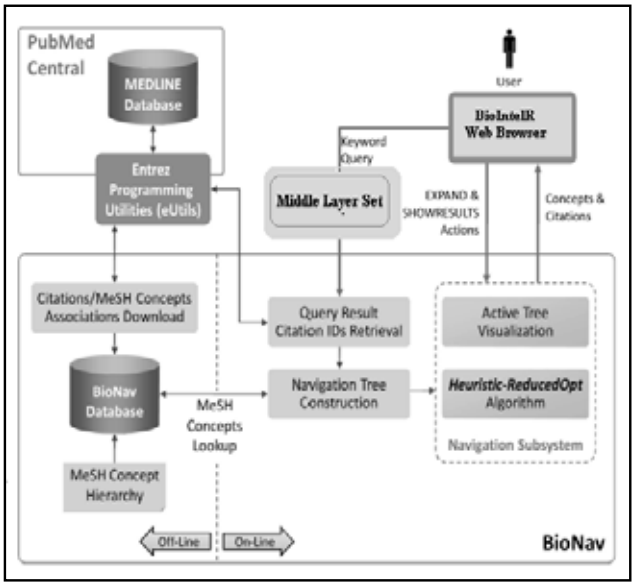

Fig. 3: BioInteiR System Architecture

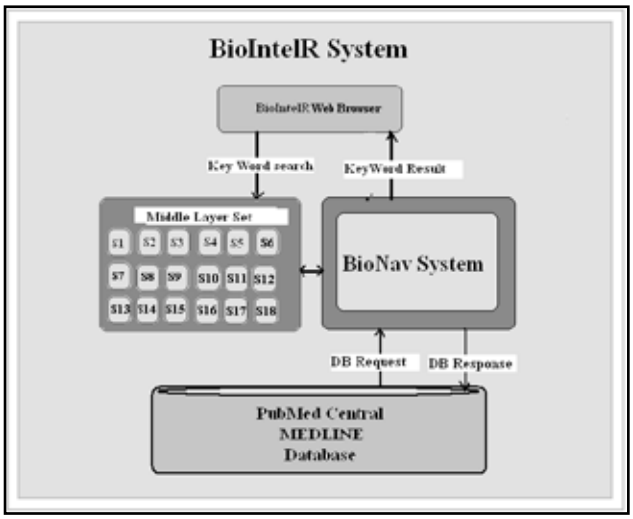

Fig. 4: BioIntelR System

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 BioIntelR 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.
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 [1], and The navigation model of BioNav which is used to device and evaluate algorithms [1].

1. 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.

2. 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.

3. 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.

4. Programming Utilities

Entrez Programming Utilities (eUtils) [7]—returns a list of citations, each associated with several MeSH concepts

5. 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 [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.

B. 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.

When a user query the database fig. 3, the middle layers specifies the correct path to run over the large central database. The proposed system adopts the BioNav system, one the path is specified and the BioNav system, The BioNav system architecture is shown in Fig. consists of two parts. The offline components populate the BioNav database with the MeSH concept hierarchy and the associations of the MEDLINE citations with MeSH concepts, while the online components support BioNav's web interface and the EXPAND-SHOWRESULTS actions of the user.

1. 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. A newly added citation may not appear immediately in the query result.

2. Online 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) [7]. 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 IðrÞ 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.

IV. 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 interactive-time 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.

A. Navigation Cost Evaluation

Fig. 5, compares the Overall navigation cost of BioIntelR over BioNav for the biochemistry query set only. BioIntelR performs better than BioNav for all queries

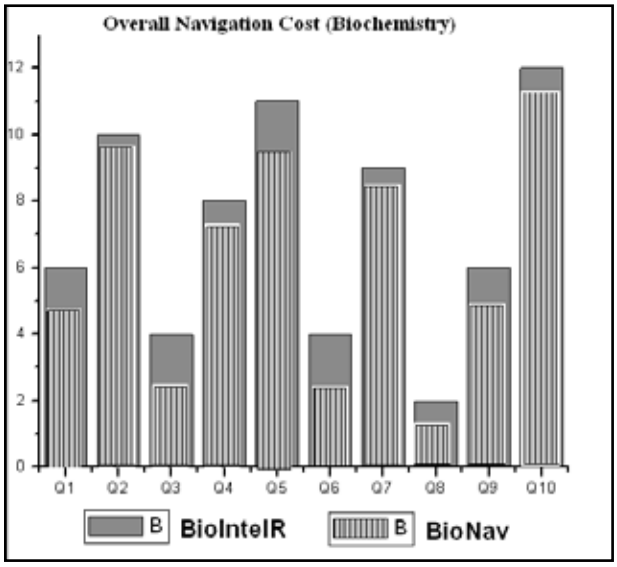

Fig. 5: Overall Navigation Cost Comparison

V. Conclusions 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, and we 552 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 4, APRIL 2011.

The implementation of BioNav is, [Online] Available: http://www.db.cse.buffalo.edu/bionav

References

[1] IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 4, April 2011.

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

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

[4] Z. Chen, 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 Infinite Expansions", pp. 176-177, Reidel, 1974.

[6] R. Delfs, A. Doms, A. Kozlenkov, 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, 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, [Online] Available: http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html, 2008.

[9] U. Feige, D. Peleg, G. Kortsarz,"The Dense k-Subgraph Problem", Algorithmica, Vol. 29, pp. 410-421, 2001.

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

[11] R. Hoffman, A. Valencia,"A Gene Network for Navigating the Literature", Nature Genetics, Vol. 36, No. 7, pp. 664, 2004.

[12] iHOP—Information Hyperlinked over Protein, [Online] Availavle: 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, 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, J. Misra,"A Linear Tree Partitioning Algorithm", SIAM J. Computing, Vol. 6, No. 1, pp. 151-154, 1977.

[16] W. Lee, L. Raschid, H. Sayyadi, 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.

[17] J. Lin, W.J. Wilbur,"Pubmed Related Articles: A Probabilistic Topic Based Model for Content Similarity", BMC Bioinformatics, Vol. 8, No. 423, 2007.

[18] D. Lindberg, B. Humphreys, A. McCray,"The Unified Medical Language System", Methods of Information in Medicine, Vol. 32, No. 4, pp. 281-291, 1993.

[19] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova,"Entrez Gene: Gene-Centered Information at NCBI", Nucleic Acids Research, Vol. 33, pp. D54-D58, Jan. 2005.

[20] Medical Subject Headings (MeSH)(2010), [Online] Available: http: //www.nlm.nih.gov/mesh/

[21] J.A. Mitchell, A.R. Aronson, J.G. Mork,"Gene Indexing: Characterization and Analysis of NLM's GeneRIFs", Proc. AMIA Ann. Symp., pp. 460-464.

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

[23] C. Perez-Iratxeta, P. Bork, M.A. Andrade,"Exploring MEDLINE Abstracts with XplorMed", Drugs of Today, Vol. 38, pp. 381-389, 2002.

[24] C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, U. Leser,"Ali Baba: PubMed as a Graph", Bioinformatics, Vol. 22, No. 19, pp. 2444-2445, 2006.

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

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

[27] H. Shatkay, R. Feldman,"Mining the Biomedical Literature in the Genomic Era: An Overview", J. Computational Biology, Vol. 10, No. 6, pp. 821-855, 2003.

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

[29] Transinsight GmbH—GoPubMed, [Online] Available: http://www.gopubmed.org/, 2008.

[30] Vivı´simo, Inc.—Clusty, [Online] Available: http://clusty.com/, 2008.

[31] XplorMed: eXploring Medline abstracts, [Online] Available: http://www.ogic.ca/projects/xplormed/, 2008.

[32] T. Zhang, R. Ramakrishnan, M. Livny,"BIRCH: An Efficient Data Clustering Method for Very Large Databases", Proc. ACM SIGMOD, pp. 103-114, 1996.

ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)



Mrs.K.RAJANI KUMARI, well known Author and excellent teacher Received M.Tech (CSE) from Andhra University, Visakhapatnam. She is working as a H.O.D in the Department of M.Tech Computer Science Engineering, Akula Sree Ramulu College Engineering & Technology, Tanuku. She has vast teaching experience in various engineering colleges. To her credit couple of publications both National and International Conferences /Journals. Her area of Interest includes Data Warehouse and Data Mining, information security, Data Communications &Networks and other advances in computer Applications. She has guided many projects for Engineering Students.



Mrs.K.S.N.V.JYOTSNA DEVI is a student of Akula Sree Ramulu college of Engineering & Technology, Tanuku. Presently she is pursuing her M.TECH (CSE) from this college and she received her Master of Computer Applications (MCA) from J.N.T University in the year 2008. She worked as an Assistant Professor in Aditya Engineering College Peddapuram. Her area of Interest includes Data Ware House and Data Mining, and Object oriented Programming Languages in computer Applications.