International Journal of Computer Applications Technology and Research
Volume 3– Issue 1, 39 - 41, 2014, ISSN: 2319–8656

Detecting and Preventing Data Overloading Difficulty using Concept Hierarchies

Vijayakumar R
Department of Computer Science and Engineering
Muthayammal Engineering College, Namakkal
Tamilnadu, India

Rajasekar G
Department of Computer Science and Engineering
Muthayammal Engineering College, Namakkal
Tamilnadu, India

Raja Rajeswari B
Department of Computer Science and Engineering
Muthayammal Engineering College, Namakkal
Tamilnadu, India

Raja S
Department of Computer Science and Engineering
Muthayammal Engineering College, Namakkal
Tamilnadu, India

Abstract: Search queries on large databases, often return a large number of results, only a small subset of which is relevant to the user. When the user want to search the result for a particular query he or she find lot of difficulties when query results are large in size. To overcome the searching and navigation difficulty the following contributions are made. First, design very good user interface to search the query using front end tools like ASP.NET and it will fetch the result from database like SQL SERVER 2005. Second, Query results are organized into a tree format using tree control. Third, Ranking concept is used to display the concepts in order based on more number of times that concept is accessed. Fourth, Edge cut algorithm is used to display the query result mostly related to the user expected results in tree format. The advantage of this proposed work is minimizing navigation cost, provided good user interface to search the query and time consuming is very less. Ranking and categorization, which can also be combined, have been proposed to alleviate this information overload problem.

Keywords: Data mining, MeSH, Edge cut algorithm, concept hierarchy, Query results.

1. INTRODUCTION

Data mining is the process of extracting knowledge from large amount of data stored in database, datawarehouse or other repositories. Concept hierarchy is defined as recursively reduce the data by replacing low level concepts(such as numeric values for age) by higher level concepts(such as young, middle-aged ,or senior) [1].

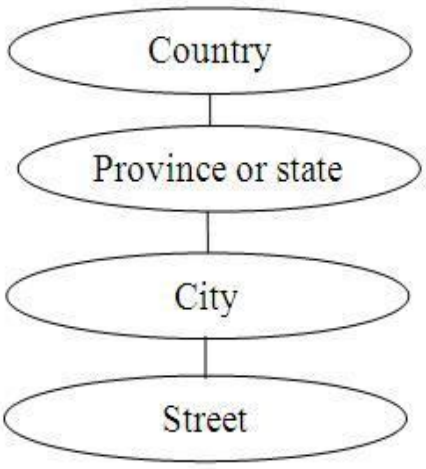

Fig: 1 Concept hierarchy

Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set. The attribute with the most distinct values is placed at the lowest level of the hierarchy. Exceptions e.g., weekday, month, quarter, year.

1.1 MeSH

Medical Subject Headings (MeSH) is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences; it can also serve as a thesaurus that facilitates searching. Created and updated by the United States National Library of Medicine (NLM), it is used by the MEDLINE/PubMed article database and by NLM's catalog of book holdings. MeSH can be browsed and downloaded free of charge on the Internet through PubMed. The yearly printed version was discontinued in 2007 and MeSH is now available online only [2] [3].Originally in English, MeSH has been translated into numerous other languages and allows retrieval of documents from different languages.

1.2 FEATURES

This MeSH mechanisms is contains following list of features.

1. MeSH is used on MEDLINE to index bibliographic citations and author abstracts from over 4,000 journals published in the United States and in 70 foreign countries (though mainly to English language papers). MEDLINE provides citations and, where available, astracts and links to full-text articles.

2. PubMed is a Web-based retrieval system developed by the National Center for Biotechnology Information

(NCBI) at the National Library of Medicine. It is part of NCBI's retrieval system called Entrez. MeSH vocabulary is used for indexing journal articles for Index Medicus® and MEDLINE and is also used for cataloging books.

3. MeSH terms are arranged in a hierarchy of "MeSH Tree Structures". When PubMed searches a MeSH term, it will automatically include narrower terms in the search, if applicable. This is also called "automatic explosion." NLM indexers examine articles and assign the most specific MeSH heading(s) that appropriately describes the concept(s) discussed. As many as 15 headings may be assigned Automatic Term Mapping feature to search for unqualified terms. When you click Go, PubMed will look for a match in up to four lists. It looks first for a match in the MeSH Translation Table. If it doesn't find a match, it looks in the Journals Translation Table, then in the Phrase List, and finally in the Author Index [2].

2. EXISTING SYSTEM

In the existing system searching is a static navigation of the database information. Here the understanding of the information is very difficult. So the categorization of the information is also a difficult task to the programmer.

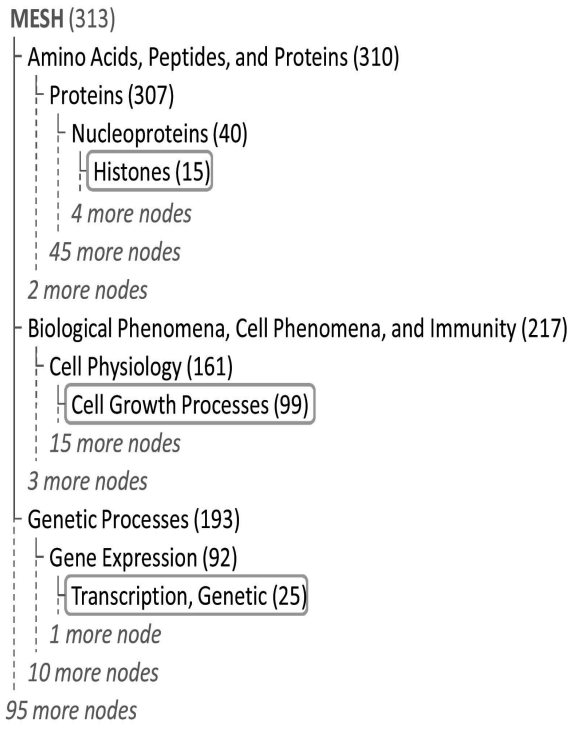

Fig 2: Static navigation on the MeSH concept hierarchy

An intuitive way to categorize the results of a query on database is by using the MeSH static concept hierarchy to build and maintain such a comprehensive structure. Each citation is associated with several MeSH concepts in two ways: 1) by being explicitly annotated with them, and 2) by mentioning those in their text.

3. PROPOSED SYSTEM

In addition to the static hierarchy navigation works mentioned above, there are works on dynamic categorization of query, which create unsupervised query-dependent results clusters, but do not study how the clusters should be navigated. When the user want to search the result for a particular query he or she find lot of difficulties when query results are large in size. The advantage of this proposed work is minimizing navigation cost, provided good user interface to search the query and time consuming is very less.

Search queries on databases, 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. To overcome the searching and navigation difficulty the following contributions are made. First, design very good user interface to search the query using front end tools like ASP.NET and it will fetch the result from database like SQL SERVER 2005. Second, Query results are organized into a tree format using tree control. Third, Ranking concept is used to display the concepts in order based on more number of times that concept is accessed. Fourth, Edge cut algorithm is used to display the query result mostly related to the user expected results in tree format.

The advantage of this proposed work is minimizing navigation cost, provided good user interface to search the query and time consuming is very less. Ranking and categorization, which can also be combined, have been proposed to alleviate this information overload problem.

4. CONCLUSION AND FUTURE WORK

Database is created to store more number of data. Front end is designed for effective navigation of query results. Tree navigation is used to display the query result in tree format. Navigation model is used to navigate the query results based on user interests. Users can expand or backtrack, based on their interests. Ranking algorithm concept is used to display the most accessed concept in first of the tree. It is very useful for every user. Best edge cut algorithm is used to get the particular set of query results from the database. Finally, this project is designed to support more number of data.

At present, ranking is based on number of times the root node is accessed. As help of these rankings, expand more number of sub-roots will be considered at future. The challenge is to implement the databases like products, college, schools and etc. Because now I have used books and authors database only.

5. REFERENCES

[1] Abhijith kashyap, Vageli Hritidis, Michalis Petropoulos, and Sotiria Tavoulari. (2011)'Effective navigation of query Results based on concept hierarchies', IEEE transaction on knowledge and engineering, VOL.23, NO.4.

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

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

[4] Chen, Z. and Li, T. (2007) 'Addressing Diverse User Preferences in SQLQuery- Result Navigation', Proc. ACM SIGMOD, pp. 641-652.

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

[6] Feige, U, Peleg, D. and Kortsarz, G. (2001),'The Dense k-Subgraph Problem', Algorithmica, vol. 29, pp. 410-421.

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

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

[9] Shatkay, H. and Feldman, R. (2003) 'Mining the Biomedical Literature in the Genomic Era: An Overview', J. Computational Biology, vol. 10,no. 6, pp. 821-855.

[10] Zhang, T, Ramakrishnan, R. and Livny, M.(1996) 'BIRCH: An Efficient Data Clustering Method for Very Large Databases', Proc. ACM SIGMOD, pp. 103-114.