Analysis of Information Retrieval using Query Extraction Techniques

[1]Tagore Kumar Tummapudi, [2]Uma M
[1]M.Tech Student, [2]Assistant Professor
SRM University, Chennai

***Abstract*— Search queries on biomedical databases, often return a large number of results, only some part of data is relevant to the user. Results categorization and ranking for biomedical databases is the focus of this work. The general way to organize biomedical data results is according to the MeSH annotations. MeSH is used by medical database for comprehensive searching of the query results. To alleviate the information overload problem Ranking and Categorization can be combined. 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. The efficiency of the results fetched by the user can be improved by using Data Mining Algorithms. Presently, few algorithms are being considered to provide the data to the fetched queries. A new approach for evaluating the relevant information for the query is done by Apriori Algorithm. Apriori gives the relevant information to the user by which the user can have easy way to access the information needed.**

***Index Terms*— MeSH, BioNav, efficient results, effectiveness**

I. INTRODUCTION

Data mining is extraction of data which is useful and it is extracted from different types of databases like biomedical, web, domain specific databases, etc. R & C for the database helps in increasing efficiency for the user to search query. Results which are displayed are irrelevant to the user. The user needs to search for the data which the user needed. The biomedical database, on which the search engine operates, contains over 18 million citations. The database is currently growing at the rate of 500,000 new citations each year [1].

The user submits an initially broad keyword- based query that typically returns a large number of results with concept hierarchies associated with MeSH concept as hierarchy. Biologists, chemists, medical and health scientists and researchers will search their data from their domain literature which need to be trustworthy. For keyword search system will use citation id which will be annotated with concept hierarchy [1].

The efficiency of the search results from the search engines varies as information providers have different levels of knowledge and different intentions. Users of query based systems are therefore confronted with the increasingly difficult task of selecting high quality information from the vast amount of web-accessible information. R & C using Apriori with the existing algorithms decreases work load. This employs the Named graphs data model for the representation of information together with quality related meta-information.

In order to optimize, the system uses concept hierarchies for navigation of query results and opt edge cut algorithm to minimize the cost and heuristic edge cut algorithm to increase the efficiency of query navigation in the biomedical database. The R & C with Apriori will improve the efficiency and effectiveness of the query results. This will increase efficiency and provide effectiveness, time saving and reduce cost of search and provides to get expected trusted results.

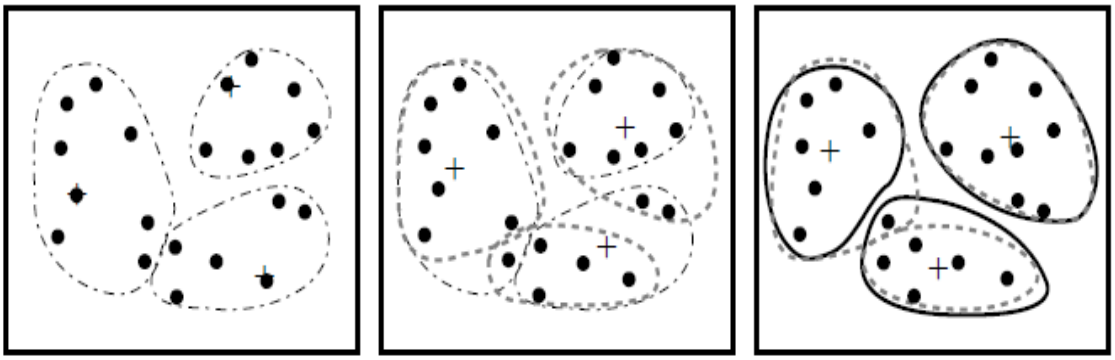
A. Optimization and categorization of Search query Results:

Optimizing of the query results fetched by the user when a query is placed on the database is done by different techniques using ranking and categorization. Ranking is done based on citation count, citation relevance, and by date. Categorization is done by using concept hierarchies, navigation tree. Navigation tree and the active tree is generated where the clustering is done and the dynamic pruning is also done to save the time and the cost of query processing. Based on k value the partitioning is done with the less number of k sub groups that is less than the more weight of the tree.

The clustering is the technique which is used in dividing the data into subsets i.e., categorizing the results. The data can be supervised or unsupervised data i.e., training set are already defined or training sets are not defined previously, for clustering or grouping of data the k value or the number of the sets should be known at the starting of the clustering or grouping is to be done.

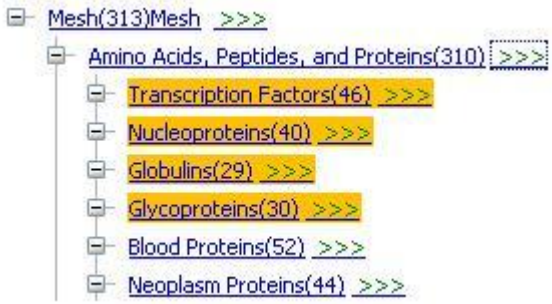
Algorithm Used
K-Means Clustering

Fig.1 K-Means Clustering



- **Data Sets**
- **Grouping of similar datasets**
- **Clusters of similar data**

Fig.2: categorization of results



The k means clustering is done based on mean distance for partitioning medians clustering is based on median distance and k medics clustering are four different techniques used for clustering the data.

B. Ranking of Search Query Results:

Ranking of the query results is done for the query results of the user. Ranking is done based on citation count, citation relevance, and by date. Ranking is done by taking the query given by the user as keyword and it verifies in the database. After verifying the database the results are retrieved.

Ranking for medical database is done by using the concept hierarchies. By using the concept hierarchies we can get the efficient results which the user wants. This reduces the information workload and navigation cost. Ranking provides the user by some relevance techniques.[7]. Ranking uses k-means algorithm to extract the top most needed or extracted, viewed lists from the web [9]. The main advantage is it increases the efficiency of extraction of the data from the web.

C. Efficiency and Accuracy for the database:

Efficiency for the search query given by the user is very necessary to maintain for the customer satisfaction. The R & C with Apriori is considered for the ranking and categorization of query results. By this R & C the user gets the query results in an accurate manner.

Efficiency and Accuracy are considered for the query results so as to decrease the information work load and navigation cost. The results are relevant for the user who searches by using the search engine. In the existing methods the user searches the query and gets lot of relevant and irrelevant data for the user. By using R&C with Apriori the user gets only the relevant data.

Algorithm Used

Apriori Algorithm Pseudo code

Procedure **Apriori** (T, *minSupport*)
{
//T is the database and *minSupport* is the minimum support
L1= {frequent items};
for (k= 2; Lk-1!=Φ; k++)
{
Ck= candidates generated from Lk-1
//that is Cartesian product Lk-1 x Lk-1 and eliminating any k-1 size itemset that is not

```
//frequent
for each transaction t in database do{
#increment the count of all candidates in Ck that are contained in t
Lk = candidates in Ck with minSupport
}//end for each
}//end for
return UkLk;
}
```

II. Conclusion and Future Work

The search results from the search engine which gives the Efficiency and accuracy for the query user searches. The search results provide the relevant data for the user. The search results from the search engines which implement the Trustworthy and High-Quality Information Retrieval System contain more accurate data with trustworthy information. Performance of retrieving trustworthy data is also improved. There are about 16 factors which affect the Content Trust of websites [8]. The future work will be providing trustworthiness which will provide high quality to the data in database.

Table 1 Comparative study of Algorithms (Expected)

	Algorithm	Processing Time (Minute)	Accuracy (%)	Unused Fields	Area Under Curve
1	K-Means	<1	80	6	0.93
2	Clustering	<2	85	3	0.75
3	Opt Edge Cut	<3	75	5	0.76
4	K-Partition	<2	80	4	0.85
5	Apriori	<1	92	8	0.95
6	SVM	<1	90	7	0.90

Presently the algorithms which are using in the medical database searching have been providing large amount of the data. By using the Apriori Algorithm we can provide the accuracy and efficiency for the query results. The Apriori uses small data sets for the information retrieval for the query results which user searches.

III. References

[1] Abhijith Kashyap, Vagelis Hristidis, Michalis Petropoulos, and Sotiria Tavoulari, "*Effective Navigation of Query Results Based on Concept Hierarchies"*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 4, APRIL 2011.

[2] Fie Wang, Noah Lee, Jianying Hu, Jimeng Sun, Shahram Ebadollahi, Andrew F.Laine ,"*A Framework For Mining Signatures From Event Sequences And its Applications In Health Care System".* IEEE Transactions on Knowledge and Data Engineering, 2013.

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

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

[5] Karthikeyan, Saravanan, Vanitha, *High Dimensional Data Clustering Using Fast Cluster Based Feature Selection*, Int. Journal of Engineering Research and Applications, ISSN: 2248-9622, Vo.4, Issue 3, March 2014, pg.65-71.

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

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

[8] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Member, and Zhaohui Zheng, *A New Algorithm for Inferring User Search Goals with Feedback Sessions*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2013.

[9] Zhixian Zhang, Kenny Q. Zhu , Haixun Wang, Hongsong Li , *Automatic Extraction of Top-k Lists from the Web*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING YEAR 2013.