ISSN:0975-9646

A Survey on Techniques to Extract High-Quality Data from the Database

MD.Khasim Pasha
M.Tech.,Computer Science and Engineering Department, SRM University, Kattankulathur, Kancheepuram Dist., Tamilnadu, India.

C.Malathy
Professor, Computer Science and Engineering Department, SRM University, Kattankulathur, Kancheepuram Dist., Tamilnadu, India.

Abstract: **Search data using queries on different type of databases like biomedical, web, domain specific databases,etc., often return a large number of results (hundreds and thousands), out of which only a small set of them are relevant to the user. By combing Ranking and Categorization search the data from the database can reduce overload. The efficiency of the results fetched by the user can be improved by enhancing trustworthy. Presently, few trust factors are being considered to provide quality data to the fetched queries. A new approach for evaluating the trust factors recommendation, user expertise, topic and incentive, provenance, decency is presented to enhance trustworthiness in this truth finder algorithm to provide trust value and quality assessment policies. Quality assessment will done to assess the quality assessment policy, provides better quality. The trustworthiness can be improved and it will provide higher quality and efficient results from the database.**
Keywords: high quality, trustworthy, trust, efficient results.

1. 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. Quality in extracted query results helps the user to fetch quality data from the database, helps in increasing efficiency of the database. Information comes from increasingly diverse sources of varying quality. There is no guarantee for the correctness of information in the database. Information quality is task-dependent. A user might consider the quality of a piece of information appropriate for one task but not sufficient for another task. Information quality is subjective, as a second less quality concerned user might consider the quality of the same piece of information appropriate for both tasks. Which quality dimensions are relevant and which levels of quality are required for each dimension is determined by the specific task .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 quality 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. The web quality assessment policy framework will enable information consumers to apply a wide range of policies to filter information. This employs the Named graphs data model for the representation of information together with quality related meta-information. The framework uses the WIQA-PL [13] policy language for expressing information filtering policies against this data model. The web search helps in the retrieval of high quality information for providing high quality data.

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 trustworthiness will improve the efficiency and effectiveness of the query results. By using trust factors recommendation, user expertise, topic and incentive, decency trustworthiness of the data will be increased. Fact confidence and website trustworthiness gives fact score which helps in evaluating trustworthiness [10]. It increases the performance in retrieving trustworthy query results. Quality is provided based on policies selected by the user by information content, contextual information and ratings of the website [9]. This will increase efficiency and provide effectiveness, time saving and reduce cost of search and provides to get expected trusted results.

A. Optimization 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. Partitioning is used for categorization and the k partitioning is the linear partitioning method that is done based on the weight of the node in the tree i.e., 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.

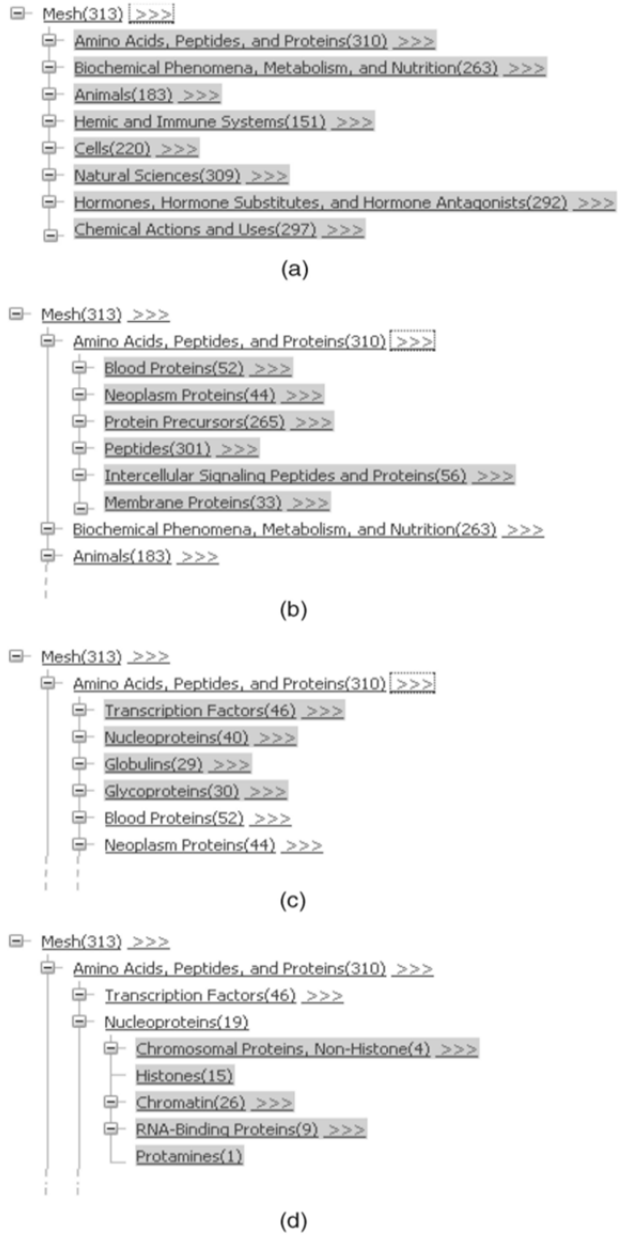

Figure 1: categorization of results

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. Quality and Trust in database:

Quality in the database is very necessary to maintain the correctness in the data in the database. The content trust is to be considered for the ranking and representation of query results. The database should always be a quality one and a correct database so that it provides higher efficiency, some amount of trust can be got by user preferred techniques but to increase the efficiency we concentrate on trust of the content and entity.

The factors that provide trust to the web content or the content in database are being calculated depending on five factors are taken to find the trust value to find the trust value based on the content trust with five factors but initially gave 0.9 for others who are not of 1 probability that give the better results and by concentrating on few other factors that increase efficiency with timely changing manner.

Quality evaluating techniques based on the content trust, information context and contextual information. These are to calculate the content trust .i.e., entity based trust. Quality can be calculated depending on trust with few more factors that may help in increasing the data quality in the database. Content trust will be providing the trust value, depending on the trust value and ratings of the content and also the context .The trust factors are recommendation, user expertise, topic and incentive provenance, decency which will help in calculating and evaluating the trust value of the query result.

Recommendation: This is about the recommendations provided by more no. of users and the experts.
User expertise: user expert in particular domain of the database.
Topic: the origin from where the data has better quality been bought.
Incentive: associations, motivations of the data are from trustable source.
Provenance: domain specific content generated from specific database.
Decency: changes with time will take place so to updated data in the database.

The trustworthiness provides the data enhanced with trust so as to provide more quality to the data. The quality assessment metrics are dependent on provenance based, user based and rating based to define the quality of the data in the database. The WIQA is to assess the quality to the data based on the context so that even the keyword can be provided with the trusted query result which will increase and enhance the quality of the data in the database and will provide the efficiency and better quality for the data in the database.

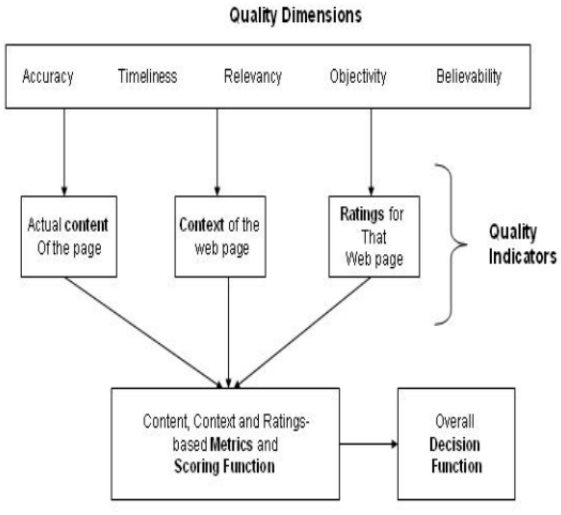

Figure 2. Architecture of Quality Module architecture

Based on the three quality indicators chosen three quality assessment metrics are classified:

1) Content based metric
2) Context based metric
3) Ratings based metric.

As Shown in figure 2 the Quality is achieved using policies selected by the user. The three quality indicators chosen are information content, contextual information and ratings of the website. Various quality dimensions like Accuracy, Timeliness, Relevancy, Objectivity, Believability etc are measured for each citation. The dimensions which are associated with each quality indicator, depends upon the policy selected by the user.

Information quality assessment metrics can be classified into three categories according to the type of information that is used as quality indicator: (1) information content itself; (2) information about the context in which information was claimed; (3) ratings about information itself or the information provider.

Context based metrics are assessed based on the provenance information taken from the metadata. Content based metrics are assessed using the information content itself. Ratings based metrics are assessed by using the rating information.

2. CONCLUSION AND FUTURE WORK:

The search results from the search engines which implement the Trustworthy and High-Quality Information Retrieval System contain more accurate data with trustworthy information. The search results provide the most truthful information. The search results are also ranked based on user-selected quality criteria. Performance of retrieving trustworthy data is also improved. There are about 16 factors which affect the Content Trust of websites [5]. The future work will be in analyzing the remaining parameters and checking their feasibility in providing trustworthiness which will provide high quality to the data in database.

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, pp. 540-553, April 2011.

[2] Karthikeyan, Saravanan, Vanitha, *"High Dimensional Data Clustering Using Fast Cluster Based Feature Selection"*, Int. Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol.4, no. 3, pg.65-71, March 2014.

[3] Christian Bizera, Richard Cyganiakb, *"Quality-driven information filtering using the WIQA policy framework"*, Web semantics: Science, Services and agents on the World Wide Web, Elsevier, vol.7, no.1, pp.1-10, January 2009.

[4] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, *Effective Pattern Discovery for Text Mining*, IEEE Transactions on knowledge and data engineering, vol. 24,no.1,pp.30-44,January 2012.

[5] Yolanda Gil, Donovan Artz, *Towards Content Trust of Web Resources*, International www conference committee, ACM, 2006.

[6] Vagelis Hristidis, Yuheng Hu, and Panagiotis G. Ipeirotis , *"Relevance-Based Retrieval on Hidden-Web Text Databases without Ranking Support"*, IEEE Transactions on knowledge and data engineering, vol. 23,no.10,pp.1555-1568 October 2011.

[7] Sumalatha Ramachandran,sujaya paulraj,sharon Joseph and vetriselvi Ramaraj, *"Enhanced Trustworthy and High Quality Information Retrieval System for Web Search Engines"*, International Journal of Computer Issues, ISSN:1694-0784,ISSN(print):1694-0841,vol. 5,pp. 38-42,2009

[8] Sana Ansari,Jayant Gadge, *"Architecture for Checking Trustworthiness of Websites"*, International journal of computer applications, ISSN (0975-8887),vol. 44,no. 14,April 2012.

[9] Xiaoxin Yin, Jiawei Han, Yu, P.S, *"Truth Discovery With Multiple Conflicting Information Providers On the Web"*, IEEE Transactions on Knowledge and Data Engineering, Vol. 20, no. 6, pp.796 – 808 , June 2008.

[10] 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, vol.25,no. 3,pp. 502-513,March 2013.

[11] 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 ,pp. 1057-1068, ISSN 1063-6382,ISBN 978-1-4673-4909-3,April 2013.

[12] http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/