ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The Efficient Prediction Review of Difficult Keyword Queries over Databases

Basim Ali Razooqi
Master of Science (Information System),
Nizam College (Autonomous),O.U,
Basheer Bagh, Hyderabad.

ABSTRACT:

Keyword queries on databases provide easy access to data, but often suffer from low ranking quality, i.e., low precision and/or recall, as shown in recent benchmarks. It would be useful to identify queries that are likely to have low ranking quality to improve the user satisfaction. For instance, the system may suggest to the user alternative queries for such hard queries. In this paper, we analyze the characteristics of hard queries and propose a novel framework to measure the degree of difficulty for a keyword query over a database, considering both the structure and the content of the database and the query results.

We evaluate our query difficulty prediction model against two effectiveness benchmarks for popular keyword search ranking methods. Our empirical results show that our model predicts the hard queries with high accuracy. Further, we present a suite of optimizations to minimize the incurred time overhead.

INTRODUCTION:

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. [1] Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining refers to the finding of relevant and useful information from data bases.

[3] Data mining or knowledge discovery is a new interdisciplinary field, merging ideas from statistics, machine learning, data bases and parallel computing. Data mining or knowledge discovery in data bases, as

it is also known, is the nontrivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analyzing changes and detecting anomalies. [7]

Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is process of finding value from volume. [3][4] Data mining in operational engineering data, manufacturers can detect faulty equipment's and determine optimal control parameters.

History of Data mining:

The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history [9]. Data mining roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning.

• Statistics are the foundation of most technologies on which data mining is built, e.g. regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals. All of these are used to study data and data relationships.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

• Machine learning is the union of statistics and AI. It could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals. Data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications.

Definition data mining:

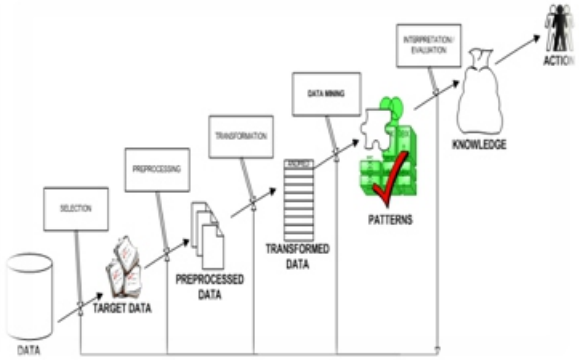


Figure (1.1): Structure of Data Mining

Data mining refers to the finding of relevant and useful information from data. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is process of finding value from volume. [2][5].

These are patterns are classified into three groups:

1.Cluster analysis:

Cluster analysis includes the groups of data records of patterns that is it grouping a set of objects. These objects are lies in same group called cluster. Cluster analysis is not algorithm but it gives solution for algorithm. Such algorithm that is clustering algorithms is based on cluster model.

2.Anomaly detection:

Anomaly detection is also known as outlier detection .It includes unusual records in data mining. It gives identification of data items.

3.Dependencies:

Dependencies include the association rule mining. It discovers relations between variables in large databases. [16].

LITERATURE SURVEY:

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system. In this paper, we analyze the features of difficult queries over databases. It proposes novel method to detect such queries. We take advantage of the structure of the data to gain insight about the degree of the difficulty of a query given the database.

"Efficient IRstyle keyword search over relational databases,"
AUTHORS: V. Hristidis, L. Gravano, and Y. Papakonstantinou

Applications in which plain text coexists with structured data are pervasive. Commercial relational database management systems (RDBMSs) generally provide querying capabilities for text attributes that incorporate state-of-the-art information retrieval (IR) relevance ranking strategies, but this search functionality requires that queries specify the exact column or columns against which a given list of keywords is to be matched. This requirement can be cumbersome and inflexible from a user perspective: good answers to a keyword query might need to be "assembled" -in perhaps unforeseen ways- by joining tuples from multiple relations. This observation has motivated recent research on free-form keyword search over RDBMSs. In this paper, we adapt IR-style document-relevance ranking strategies to the problem of processing free-form keyword queries over RDBMSs.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Our query model can handle queries with both AND and OR semantics, and exploits the sophisticated single-column text-search functionality often available in commercial RDBMSs.

"SPARK: Top-k keyword query in relational databases,"
AUTHORS:Y. Luo, X. Lin, W. Wang, and X. Zhou

With the increasing amount of text data stored in relational databases, there is a demand for RDBMS to support keyword queries over text data. As a search result is often assembled from multiple relational tables, traditional IR-style ranking and query evaluation methods cannot be applied directly.

In this paper, we study the effectiveness and the efficiency issues of answering top-k keyword query in relational database systems. We propose a new ranking formula by adapting existing IR techniques based on a natural notion of virtual document. Compared with previous approaches, our new ranking method is simple yet effective, and agrees with human perceptions.

PROBLEM STATEMENT:

There have been collaborative efforts to provide standard benchmarks and evaluation platforms for keyword search methods over databases. One effort is the data-centric tracks of INEX Workshop Queries were provided by participants of the workshop. Another effort is the series of Semantic Search

Challenges (SemSearch).The results indicate that even with structured data, finding the desired answers to keyword queries is still a hard task.more interestingly, looking closer to the ranking quality of the best performing methods on both workshops.

Disadvantages of Existing System:

» Suffer from low ranking quality.

» Performing very poorly on a subset of queries.

PROPOSED SYSTEM:

» We set forth a principled framework and proposed novel algorithms to measure the degree of the difficulty of a query over a DB, using the ranking robustness principle.

» Based on our framework, we propose novel algorithms that efficiently predict the effectiveness of a keyword query.

Advantages:

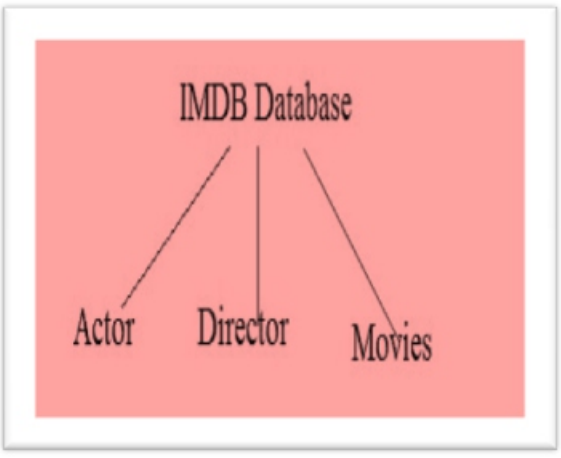
» Easily mapped to both XML and relational data.

» Higher prediction accuracy and minimize the incurred time overhead.

IMPLEMENTATION:

Data and Query Modeling:

In this module, first we develop a System Model for our proposed System. We model a database as a set of entity sets. Each entity set S is a collection of entities E. For instance, movies and people are two entity sets in IMDB Keyword query interface (KQIs) evaluated on well-known IMDB (Intermediate Database) data set. This data set contains structured information about movies and people in given business. IMDB database mainly contain three tables that is actor, director and movies [3] [5].



Figuer(5.1): Intermediate Database(IMDB)

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Methods uses for performance of information retrieval system are measure in relevant document and non-relevant document .These are having terms as follows:

- **Precision:**

It is the fraction of documents retrieved that are relevant to user's information need.

- **Recall:**

It is the fraction of the documents that are relevant to the query that are successfully retrieved to user.

• Average Precision: The results show that our method predicts the degree of the difficulty of a query efficiently and effectively.[4]

Pros:

1. Keyword search provides user friendly interface rather than X path and X query.

2. XML is used to store data in XML document format rather than table format.

3. XML provide security to data, user not easily recognize the XML data rather than traditional Table format.

4. User enters a keyword (i.e. Attribute, key, identifier).

5. SR scores measures the difficulty of queries over database.

6. The algorithm to compute the SR score, and parameters to tune its performance.

7. It gives small time overhead compared to the query execution time.

We ignore the physical representation of data in this paper. That is, an entity could be stored in an XML file or a set of normalized relational tables.

The above model has been widely used in works on entity search and data-centric XML retrieval, and has the advantage that it can be easily mapped to both XML and relational data.

Keyword query interfaces:

Keyword query interfaces (KQIs) for databases provides flexibility and ease of use in searching and exploring data [1][5]. Keyword queries have potential answer in data set. Keyword query interfaces identifies the information needs or requirements behind every keyword query. It rank the answer so it appears at the top of the list [1][6]. Database is the collection of data and this data is organizes to model used to support internal operations. In this a computer program quickly selects pieces of data. Databases contain entities, and entities contain attributes that take attribute values. Some of the difficulties of answering a query are as follows:

First:

Unlike queries in languages like SQL, users do not normally specify the desired schema element(s) for each query term. For instance, query Q1:Godfather on the IMDB database does not specify if the user is interested in movies whose title is Godfather or movies distributed by the Godfather Company. Thus, a KQI must find the desired attributes associated with each term in the query.

Second:

The schema of the output is not specified (i.e., users do not give enough information to single out exactly their desired entities).

For example:

Q1 may return movies or actors or producers. We present a more complete analysis of the sources of difficulty and ambiguity in Section Recently, there have been collaborative efforts to provide standard benchmarks and evaluation platforms for keyword search methods over databases. One effort is the data centric track of INEX Workshop where KQIs are evaluated over the well-known IMDB data set that contains.

Ranking for Structured Data:

In this module we present the Ranking Robustness Principle, which argues that there is a (negative) correlation between the difficulty of a query and its ranking robustness in the presence of noise in the data.The degree of the difficulty of a query is positively correlated with the robustness of its ranking over the original and the corrupted versions of the collection. We call this observation the Ranking Robustness Principle [13].

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Corruption Module:

The first challenge in using the Ranking Robustness Principle for databases is to define data corruption for structured data. For that, we model a database DB using a generative probabilistic model based on its building blocks, which are terms, attribute values, attributes, and entity sets.

A corrupted version of DB can be seen as a random sample of such a

Probabilistic model. which is already computed

Ranking Module:

Each ranking algorithm uses some statistics about query terms or attributes values over the whole content of DB. Some examples of such statistics are the number of occurrences of a query term in all attributes values of the DB or total number of attribute values in each attribute and entity set. These global statistics are stored in M (metadata) and I (inverted indexes) in the SR Algorithm pseudo code [6].

CONCLUSION:

We introduced the novel problem of predicting the effectiveness of keyword queries over DBs. We showed that the current prediction methods for queries over unstructured data sources cannot be effectively used to solve this problem. We set forth a principled framework and proposed novel algorithms to measure the degree of the difficulty of a query over a DB, using the ranking robustness principle. Based on our framework, we propose novel algorithms that efficiently predict the effectiveness of a keyword query. Our extensive experiments show that the algorithms predict the difficulty of a query with relatively low errors and negligible time overheads.

REFERENCES:

[1] V. Hristidis, L. Gravano, and Y. Papakonstantinou, "Efficient IRstylekeyword search over relational databases," in Proc. 29thVLDB Conf., Berlin, Germany, 2003, pp. 850–861.

[2] Y. Luo, X. Lin, W. Wang, and X. Zhou, "SPARK: Top-k keywordquery in relational databases," in Proc. 2007 ACM SIGMOD,Beijing, China, pp. 115–126.

[3] V. Ganti, Y. He, and D. Xin, "Keyword++: A framework toimprove keyword search over entity databases," in Proc. VLDBEndowment, Singapore, Sept. 2010, vol. 3, no. 1–2, pp. 711–722.

[4] J. Kim, X. Xue, and B. Croft, "A probabilistic retrieval modelfor semistructured data," in Proc. ECIR, Tolouse, France, 2009,pp. 228–239.

[5] N. Sarkas, S. Paparizos, and P. Tsaparas, "Structured annotationsof web queries," in Proc. 2010 ACM SIGMOD Int. Conf. Manage.Data, Indianapolis, IN, USA, pp. 771–782.

[6] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, andS. Sudarshan, "Keyword searching and browsing in databasesusing BANKS," in Proc. 18th ICDE, San Jose, CA, USA, 2002,pp. 431–440.

[7] C. Manning, P. Raghavan, and H. Schütze, An Introductionto Information Retrieval. New York, NY: Cambridge UniversityPress, 2008.

[8] A. Trotman and Q. Wang, "Overview of the INEX 2010 datacentric track," in 9th Int. Workshop INEX 2010, Vugh, TheNetherlands, pp. 1–32,

[9] T. Tran, P. Mika, H. Wang, and M. Grobelnik, "Semsearch ´S10,"in Proc. 3rd Int. WWW Conf., Raleigh, NC, USA, 2010.

[10] S. C. Townsend, Y. Zhou, and B. Croft, "Predicting query performance,"in Proc. SIGIR '02, Tampere, Finland, pp. 299–306.

AUTHORS BIOGRAPHY:

BASIM ALI RAZOOQI ,pursuing his Master of Science in Information System, from Nizam College (Autonomous),O.U,Basheer Bagh, Hyderabad,India.