Novel Algorithm for Faster Multi-Keyword Search in Structured Databases

Abhishek Kumar Dinkar[1], Sayar Singh Shekhawat[2]
[1]*M. Tech. Scholar, Department of Computer Science, Arya Institute of Engineering & Technology, Jaipur, Rajasthan*
[2] *Head of Department, Department of Computer Science, Arya Institute of Engineering & Technology, Jaipur, Rajasthan*

***Abstract*- Searching the major part of the database operation. For accessing any of the information from the database it is always required to perform the search. If the search process is efficient then the time taken the get the required information will also be less. In the scenario which we have taken contain the information organized in the structured data for the big organization like any automobile industry maintaining information regarding its department. Now if we perform the keyword based query, then on based of the keyword the queries will be formed.**
So, in order to reduce the time involved in the formation of the queries from the table we have suggested the use of the associative mapping table in the search mechanism which will reduce the time involved in the process

***Index Terms*- Database, Keyword Search, Structured Organization of Data.**

1. INTRODUCTION

Big data is set to offer companies tremendous insight. But with terabytes and petabytes of data pouring into organizations today, traditional architectures and infrastructures are not up to the challenge. IT teams are burdened with ever-growing requests for data, ad hoc analyses and one-off reports. Decision makers become frustrated because it takes hours or days to get answers to questions, if at all. More users are expecting self-service access to information in a form they can easily understand and share with others.

In today's hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly. Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed. The challenge only grows as the degree of granularity increases. One possible solution is hardware. Some vendors are using increased memory and powerful parallel processing to crunch large volumes of data extremely quickly. Another method is putting data in-memory but using a grid computing approach, where many machines are used to solve a problem. Both approaches allow organizations to explore huge data volumes and gain business insights in near-real time.

Even if you can find and analyze data quickly and put it in the proper context for the audience that will be consuming the information, the value of data for decision-making purposes will be jeopardized if the data is not accurate or timely. This is a challenge with any data analysis, but when considering the volumes of information involved in big data projects, it becomes even more pronounced. Again, data visualization will only prove to be a valuable tool if the data quality is assured. To address this issue, companies need to have a data governance or information management process in place to ensure the data is clean. It's always best to have a pro-active method to address data quality issues so problems won't arise later.

Different communities are working on the Big Data challenge. For example, the systems community is developing technologies for massive storage of big data. The network community is developing solutions for managing very large networked data. The data community is developing solutions for efficiently man-aging and analyzing large sets of data. Big Data research and development is being carried out both in

academia, industry and government research labs. However, little attention has been given to security and privacy considerations for Big Data. Security cuts across multiple areas including systems, data and net-works. We need the multiple communities to come together to develop solutions for Big Data security and privacy.

Keyword Query System Model

This section describes, in general terms, the framework of a keyword query system[7]. A keyword query system is a complex system which is capable of taking as input, a set of words, called keywords and give an appropriate answer. Here, the structure and semantics of the answer is specific to the query answer system. A system for keyword query consists of the following:

(i) Data Model: It describes the high-level representation of the data in the system, such that it reflects the constraints, associations, and organization of the data. The actual implementation of the representation is not of concern, here.

(ii) Query Model: It specifies the structure of the input that can be given to the system. For keyword queries, the most common form of input is a set of words or terms. This simplifies the task of querying, since, the user is required to know neither any query language nor the schema of the database. A more powerful form of querying is by using graph or tree patterns.

(iii) Answer Model: It specifies the structure of an answer to a query and the requirements that it must satisfy according to the semantics of the system. The answers are usually represented as a graph or tree, or it may be just a tuple or a term.

(iv) Scoring Model: In general, there will be many answers to the same keyword query and hence most of the systems employ a scoring model, which assigns a score to each of the answers, based on their relevance. The notion of relevance is very ambiguous, since it depends on the user. Also, since keyword querying is not a powerful method in terms of expressiveness, the users are unable to articulate their requirements exactly. Hence, instead of a single answer, the system must return top few documents with the highest scores. The score is dependent on the semantics of the system. A simple method used, is to give higher score to an answer with smaller number of joins. But, most systems use complex rules to assign scores, to improve the quality of the top ranked answers.

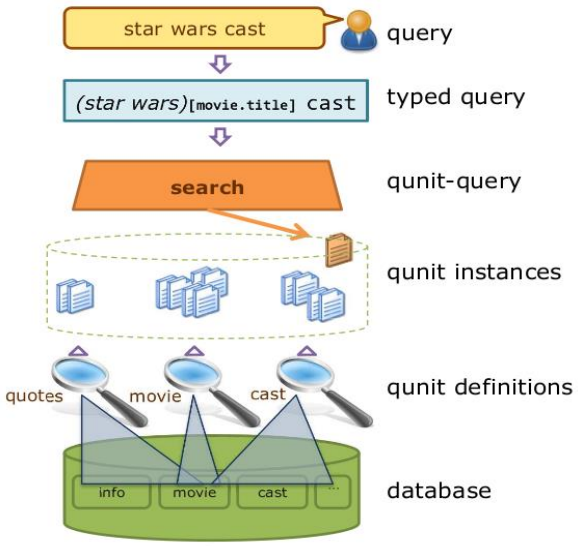

Fig 1 Searching Concepts in Database

2. LITERATURE REVIEW

Sonia Bergamaschi Elton Domnori Francesco Guerra [1], Keyword queries offer a convenient alternative to traditional SQL in querying relational databases with large, often unknown, schemas and instances. The challenge in anstheyring such queries is to discover their intended semantics, construct the SQL queries that describe them and used them to retrieve the respective tuples. Existing approaches typically rely on indices built a-priori on the database content. This seriously limits their applicability if a-priori access to the database content is not possible. Examples include the on-line databases accessed through theyb interface, or the sources in information integration systems that operate behind wrappers with specific query capabilities.

According to paper "Big Data: Review, Classification and Analysis Survey K.Arun, Dr. L.Jabasheela"[2]. International Journal of Innovative Research in Information Security (IJIRIS).This paper the author has given the information regarding the big data and impact on the World Wide Web. This paper also describes about the advent of Big Data, its Architecture and Characteristics. Here author discussed about the classifications of Big Data to the business needs and how for it will help us in decision making in the business environment.

Another paper is "Big-Data Security, KalyaniShirudkar, DilipMotwani, International Journal of Advanced Research in Computer Science and Software Engineering"[3]. In this paper author has presented challenges in the field of "Big data Security". Big data suffers from number of challenges which are related to security like-computation in distributed programming, security of data storage and transaction log, input filtering from client, scalable data mining and analytics, access control and secure communication.

Another paper is "A Review on Cloud to Handle and Process Big Data" Nishu Arora, Rajesh Kumar Bawa, International Journal of Innovations & Advancement in Computer Science IJ IACS[4]. The idea behind the Cloud computing is that user can use the service anytime, anywhere through the Internet, directly through the use of the browser. And in the field of the cloud computing data is saved in virtual space as it uses the browsers to use network services. And also the networks are associated so main concern is regarding the security of data. In this paper the author has described about the big data , its types and how it can be useful with the help of cloud computing.

Another paper is "NoSQL Database: New Era of Databases for Big data Analytics-Classification, Characteristics and Comparison, By A B M Moniruzzaman, Syed Akhter Hossain" International Journal of Database Theory and Application[5]. NoSQL, for "Not Only SQL," which relates group of non-relational data management systems; where databases are not built mailnly on tables, and also do not use SQL for data manipulation. NoSQL database management systems are also beneficial when working with a huge quantity of data when the data's nature does not require a relational model.

Another paper is "BIG Data and Methodology-A review, By Shilpa, Manjit Kaur Volume 3, Issue 10, October 2013"[6].In this article, the author given an overview about the big data's concept , its tools, its techniques, its applications, advantages and challenges have been reviewed. And the results have given away that regardless of the fact that accessible information, tools and techniques available in the literature, there are numerous focuses to be viewed as, discussed, analyzed, developed, and improved, and so on.

ErolGelenbe and Omer H. Abdelrahman [7] proposed that the searching the Internet for some object characterized by its attributes in the form of data, such as a hotel in a certain city whose price is loTheyr than some amount, is one of our most common activities when They access the Theyb. They discuss this problem in a general setting, and compute the average amount of time and energy it takes to find an object in an infinitely large search space.

ShengliWu,ChunlanHuang,Jieyu Li [8] , proposed that for information retrieval systems, the collection of documents becomes larger and larger. For some query, an information retrieval system needs to retrieve a large number of documents as the result to the query. In reality, very often people mainly care about some top-ranked documents rather than the complete long list of documents. In such a situation, how to develop a retrieval system with desirable efficiency and effectiveness is a research problem. In this paper, they focus on the data fusion approach to information retrieval, in which each component retrieval system contributes a result and all the results are combined by a combination method. The goal of this research is to find a feasible combination method that is able to balance effectiveness and efficiency. Using 3 groups of historical runs from TREC for the experiment, they find that with the weights trained by weighted linear regression, the linear combination method can achieve good results in effectiveness and efficiency.

3. PROPOSED WORK

In our proposed approach we have make use of two algorithms,

Algorithm 1: For Keyword Search

Step1: Capture the Keyword String user entered for Searching

Step 2: Split the multi-keyword string into an array. Now each element of array is the keyword to be searched.

Step 3: In the keyword search, we will maintain the following data structures,

Structure 1:

Table name
Fieldname
Keyword matched

By making this structure we will get access the table name and fieldname of the table containing the keyword.

Step 4: If the keyword search in not found then the statistical analysis of the counting the occurrence of the keyword in the columns of the table is performed and then the information is transferred to the table with structure 1.

Step 5: If the search keyword is found in the structure 1 , then the navigation in the tables is not required so the quick access to the related table query will be achieved and thus faster access of the data.

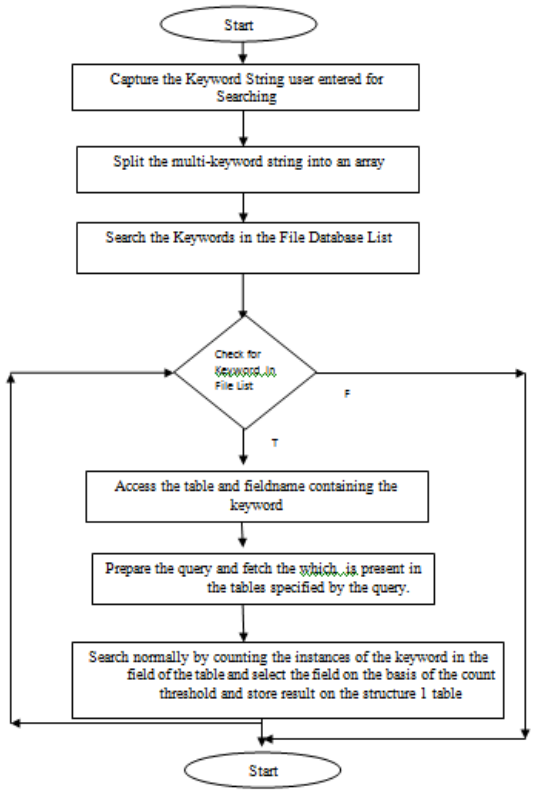

Figure 2. Proposed Algorithm

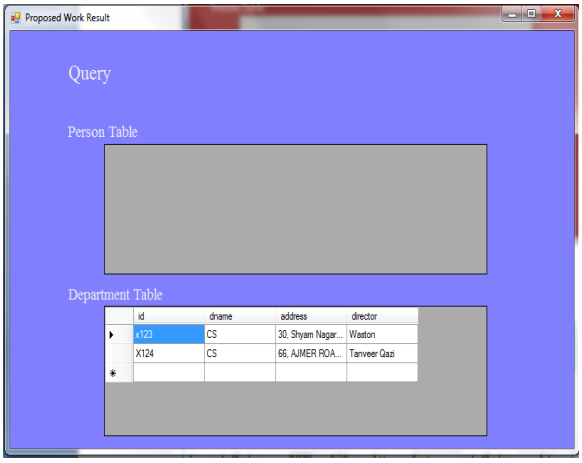

Figure 3. Proposed Implementation

This form is used for specifying the keyword which we want to search in the database file. The keyword can be a single string or multiword string.

In the Figure 3 the proposed implementation for the query formation is shown , in this the query is form using following structure ,

Structure 1:

Table name
Fieldname
Keyword matched

Perform the search and quickly fetch the matched fieldname and table name and if the keyword is not found in the structure 1 table then the algorithm of the base paper, in which we list all the fields of all the tables and on the basis of the frequency of occurrences of keywords in the columns , the columns are selected after comparing the count with the threshold value.

Test Results

Case I Single Word Search

The single word "CS" is searched , then the Base works and proposed comparison is show in the table 1

Base Work	Proposed Work
2612 ms	131 ms

Table.1 Keyword "CS" search compared in Base as well as proposed Implementation

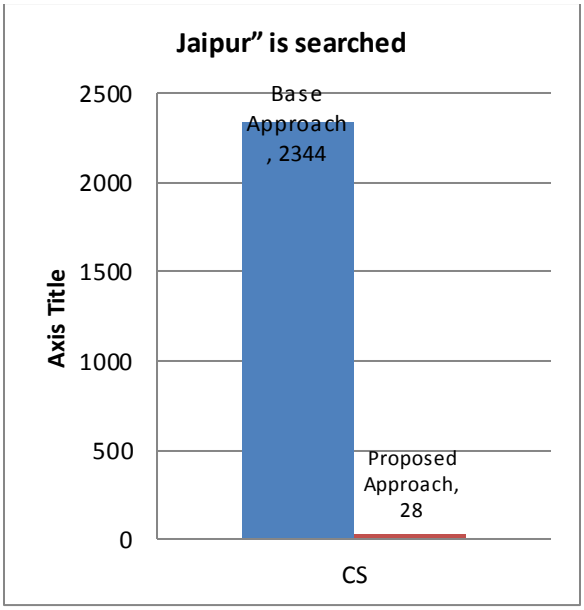

Figure 4. Comparison for Keyword CS

The single word "Jaipur" is searched, then the Base works and proposed comparison is show in the table 2

Base Work	Proposed Work
2344 ms	28 ms

Table 2 Keyword "Jaipur" search compared in Base as well as proposed Implementation

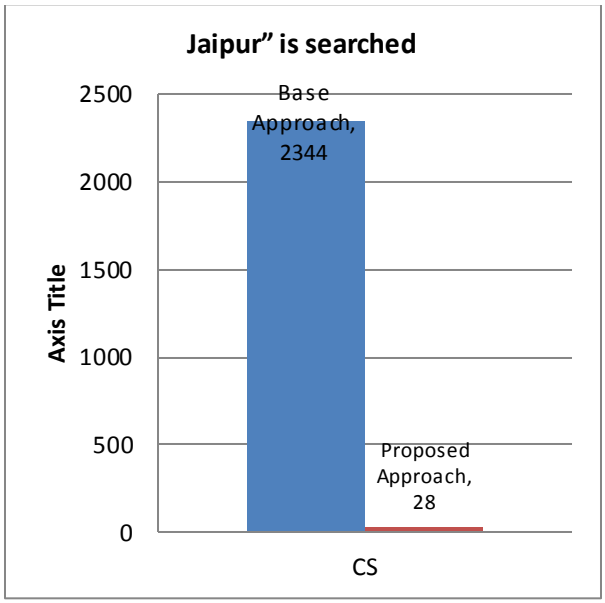

Figure 5. Comparison for Keyword Jaipur

Case II Multiple Word Search

The single word "CS Jaipur" is searched , then the Base works and proposed comparison is show in the table 4.1

Base Work	Proposed Work
2967 ms	231 ms

Table 3 Keyword "CS Jaipur" search compared in Base as well as proposed Implementation

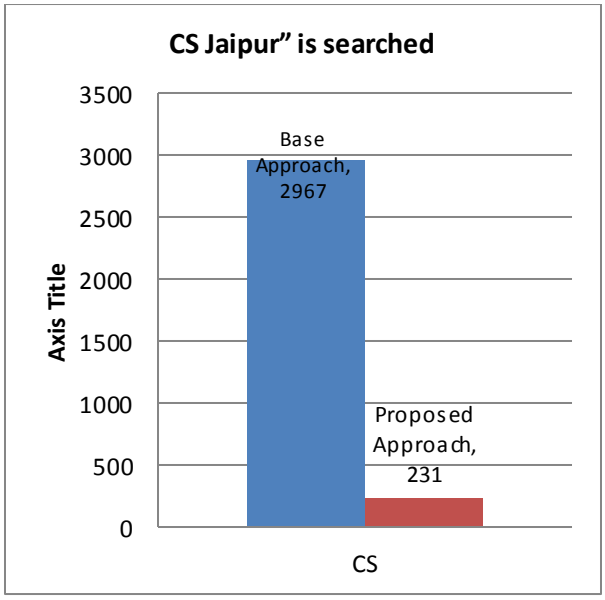

Figure 5. Comparison for Keyword CS Jaipur

4. CONCLUSION

The big data is a very vast field and much work is required to be done in this field and there is always requirement for the better algorithm for the searching of any information in the big data and in the future work we will try to extend our research on the cloud computing and parallel processing of the big data.

REFERENCES

[1] Sonia Bergamaschi Elton Domnori Francesco Guerra.Keyword Search over Relational Databases: A Metadata Approach,2011.

[2] KalyaniShirudkar, DilipMotwani "Big-Data Security" International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 3, March 2015.

[3] Nishu Arora, Rajesh Kumar Bawa "A Review on Cloud to Handle and Process Big Data" International Journal of Innovations & Advancement in Computer Science IJ IACS ISSN 2347 – 8616 Volume 3, Issue 5 July 2014

[4] A B M Moniruzzaman, Syed Akhter Hossain "NoSQL Database: New Era of Databases for Big data Analytics-Classification, Characteristics and Comparison" International Journal of Database Theory and Application Vol. 6, No. 4, August, 2013.

[5] Shilpa, Manjit Kaur "BIG Data and Methodology-A review" Volume 3, Issue 10, October 2013 ISSN: 2277 128X.

[6] Gary Pan, Seow Poh Sun, Calvin Chan and Lim Chu Yeong,"Analytics and Cybersecurity: The shape of things to come",CPA ,2015

[7] Erol Gelenbe and Omer H. Abdelrahman,"Search in the Universe of Big Networks and Data",IEEE ,2014

[8] Shengli Wu,Chunlan Huang,Jieyu Li,"Combining Retrieval Results for Balanced Effectiveness and Efficiency in the Big Data Search Environment",IEEE International Conference on Computer and Information Technology,2014

[9] Ajeet Lakhani,Ashish Gupta,K. Chandrasekaran,"IntelliSearch: A Search Engine based on Big Data Analytics integrated with Crowdsourcing and category-based

search",International Conference on Circuit, Power and Computing Technologies ,2015

[10] Zhihua Xia, Member, Xinhui Wang, Xingming Sun, and Qian Wang,"A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data",IEEE,2015

[11] Bing Wang, Wei Song, Wenjing Lou Y. ,Thomas Hou,"Inverted Index Based Multi-Keyword Public-key Searchable Encryption with Strong Privacy Guarantee",IEEE Conference on Computer Communications (INFOCOM),2015.