Scientific Journal of Impact Factor (SJIF) : 3.134

ISSN (Print) : 2348-6406
ISSN (Online): 2348-4470

International Journal of Advance Engineering and Research Development

Volume 1, Issue 11, November -2014

PERFORMANCE ANALYSIS OF KEYWORD SEARCH ENGINE

Ms. Uma Nalawade[1], Mrs. Kanchan Varape[2]

Computer Engineering, JSPM's NTC Pune, uma.nalawade300190@gmail.com
Computer Engineering, JSPM's NTC Pune, kanchanv2007@gmail.com

Abstract- *In past years the researches on the keyword search engine have increased and now it has became an active area for research in relational database and Information retrieval (IR). Till today so many this are been proposed and implemented too but still there is missing of standardizations. Due to this lack of standardization there are different evaluations and resultant information is present. In this survey the comparison of the already proven theories is done and the performance evaluation of relational keyword search systems is also done with the help of various theories. After comparing the various theories , memory consumption precludes many search techniques from scaling beyond small datasets with tens of thousands of vertices. We also explore the relationship between execution time and factors varied in previous evaluations; our analysis indicates that these factors have relatively little impact on performance. At last we have shown the unacceptable performance of the systems and underscores the need for standardization which is implied by the community of information retrieval system evaluating these retrieval system.*

Keywords- *Information retrieval, search engine, performance analysis, structured data, unstructured data*

I. INTRODUCTION

The keyword search is the most popular information search methodology which gives the provision to user that it is not necessary to know the structure of data as well as the query language. The search engine or tool present in market are nothing but the boon for the user who does not know the basics of the data as well as what exactly the query language is. In search engines we can see the text box is provided for the user input at the top of all the documents. There are many searching techniques or the information retrieval methods present in market and off course currently used by all levels of users. Most of the methods claim that they provide best performance but there is no equality in the performance. There are many theories are declared and implemented too but still there is lack of standardization. This survey will focuses on the different methodologies proved by some people and their opinion about these overall searching techniques. Paper will tell how what is overview of the data i.e. the structured and unstructured data, how the keyword search is been implemented and lastly it will say about the overall characteristics of the searching and information retrieval. In this paper section I will gives the brief introduction about "how the information retrieval" has became the active searching topic and what are the different factors affecting to it. Section II will give brief introduction about the characteristics of the keyword search, section III of this paper will give the genesis of information retrieval and some characteristic, section VI will brief about the different measures of the information searching and retrieving, and finally the section V will brief about the effective measures and section VI will compare the resultant content and conclude. The overall paper talks about the different information retrieval techniques and the comparisons made in between proved theories.

II. OVERVIEW

Now days the web use is been increased plenty and also the users area unit pretty much increased . The survey provides the note that though' the utilization of net is increased still the performance isn't that abundant improved. Additionally the paper tells that the performance analysis of the search engines has become the present active topic. On this subject several theories with their implementation area unit gift however there's lack of standardization here. The performance of any quite search depends upon the question and therefore the quite information i.e. whether or not the info is structured or unstructured. These two things area unit greatly vital. Principally in programs the hierarchal keywords area unit used for raising the general performance of that search engine. Here we tend to area unit planning to see one by one however the performance of data retrieval varies. With the number of obtainable text information in relative databases growing speedily, the necessity for standard users to look such data is dramatically increasing. Although the foremost RDBMSs have provided full-text search capabilities, they still need users to own data of the info schemas and use a structured command language to look data. Though the increasing speed of information retrieval is currently very much active topic there is lack of standardization because of the varying results. There are so many proven theories as well as implemented too still there is no any standard for a particular section in information retrieval.

@IJAERD-2014, All rights Reserved

The amount of accessible structured information (in net or computer network or perhaps on personal desktops) for normal users grows chop-chop. Besides information varieties like range, date and time, structured information bases typically additionally contain an oversized quantity of text data, like names of individuals, organizations and merchandise, titles of books, songs and films, street addresses, descriptions or reviews of merchandise, contents of papers, and lyrics of songs, etc. the necessity for normal users to seek out info from text in these databases is dramatically increasing. Applying the keyword search techniques in text databases (IR) to relative databases (DB) may be a difficult task as a result of the 2 styles of databases area unit totally different. First, in text databases, the fundamental info units searched by users area unit documents. For a given keyword question, IR systems figure a numeric score for every document and rank the documents by this score. the highest graded documents area unit came as answers. In relative databases, however, info is hold on within the type of columns, tables and first key to foreign key relationships. The logical unit of answers required by users isn't restricted to a personal column price or perhaps a personal tuple; it should be multiple tuples joined along. the normal search model in relative information needs users to possess data of the database schema and to use a structured search language like SQL or QBE-based interfaces. albeit most of the foremost RDBMSs have integrated full-text search capabilities victimization relevance-based ranking ways developed in info retrieval (IR), they still have the on top of 2 needs for users.

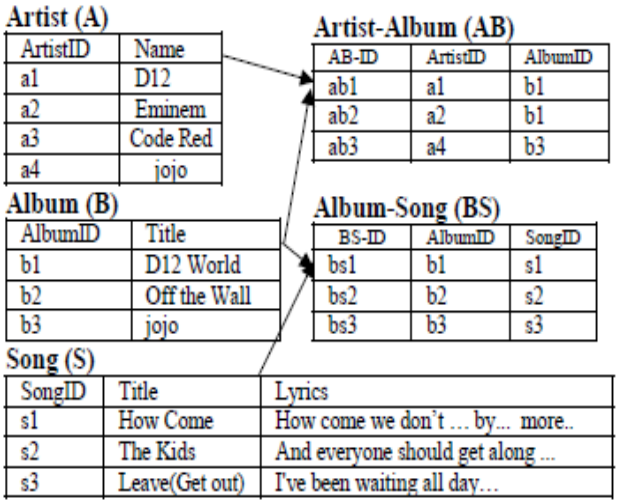

***Figure 1. Lyrics Database Example*[5]**

The need for standard users to search out data from text in these databases is dramatically increasing. the target of this paper is to produce effective search of text data in relative databases. we tend to take a lyrics info (Figure 1) as AN example as an instance the matter. There are five tables within the lyrics info. Table creative person has one text column: Name. Table Album has one text column: Title. Table Song has two text columns: Title and Lyrics. The tuples of Table creative person and people of Table Album have m:n relationships (an album could also be created by multiple creative persons and an artist might manufacture quite one album), and Table Artist- Album is that the corresponding relationship table. Table Song-Album is additionally a relationship table capturing the m:n relationships between tuples of Album and Song[5]. As shown in the figure 1 the ArtistID can be a keyword and by using it we can retrieve the songs sung by that particular singer, this could be more specific if we add name of album in search. In this way the query will become more specific and of course the result which will be displayed that will be more correct or approximately good

.

III. CHARACTERISTICS OF KEYWORD SEARCH

Users perform searches to satisfy data wants. A keyword question is associate degree expression of such associate degree data want, and it's the task of the retrieval system to come back data things that area unit relevant thereto want. For unstructured text, the data things area unit distinct documents. For relative knowledge, however, the data things area unit (possibly joined) tuples. The relative search system thus has the extra responsibility of deciding the candidate tuple joins. in addition, the keyword question contains no schema data, in order that every keyword probably should be matched against every field of the joined tuple. In a very structured source language like SQL, there's just one correct answer set. In distinction, there square measure several plausible similarity metrics, every with its own manner of inferring a user's data want from a question , and of hard the query's similarity to data things, to get a ranking of answers. The effectiveness of a

@IJAERD-2014, All rights Reserved

response to a keyword question, and therefore of the similarity metric, isn't one thing that may be formally proved; rather, it's determined by the user World Health Organization complete the data want, developed the question, and perused the response. This effectiveness should be by trial and error assessed.

The keywords are nothing but used in all kinds of databases for instance we can see in indexing and hashing techniques the index will be set. This index will be helpful at the time of searching in database. Here we can also consider the example of library managing tool in which the books can be searched by using author names or a particular subject like software engineering, so here in this case the name of author or the subject name will automatically become the search key. Because of these keywords the search can become faster and specific. The structured and unstructured data can be analyzed with the help of these kind of keywords and the process of information retrieval can become more faster and also much easier for the non-computer background person.

IV. EVALUATION IN INFORMATION RETRIEVAL

The empirical approach to data retrieval began with the sphere itself, within the experiments conducted at the library of the Cranfield physical science school, England, within the late Nineteen Fifties and early Sixties, below the direction of the bibliotheca, Cyril Cleverdon. The orthodoxy of the time in IP was that advanced, class-conscious classification schemes were essential to effective retrieval. The question launched to answer was that classification theme was best; and also the answer the experiments got hold of was none. It created no nice distinction that theme was used; merely classification documents by plain keywords was pretty much as good a technique as any; what mattered was the method of retrieval. Cleverdon himself represented these as "results that appear to offend against each canon on that we have a tendency to be trained as librarians". The process of classification will be also useful methodology to improve the overall performance of search engines. As the various proved theories talk about the performance improvement techniques, so this classification of data into similar categories can also be a good option. There are huge databases like bank, companies etc. these databases are always use the classification methodologies so that the maintenance and retrieval of the information could be more easier as well as faster.

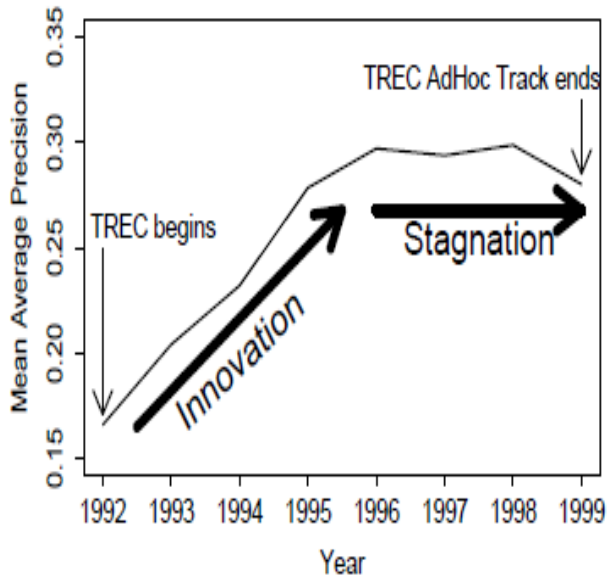

***Figure 2. Retrieval Effectiveness Of The SMART Version From The First Eight Years Of TREC*[7]**

The Cranfield experiments themselves were meted out entirely, and rather heroically, by manual means; the two hours needed to method every of the 361 searches by hand was thought to be "relatively low cost compared to what would are the value for any style of machine searches" [7]. The prevalence of free text search in internet search engines has impressed recent interest in keyword search on relative databases. Whereas relative queries formally specify matching tuples, keyword queries area unit general expressions of the user's info want. The correctness of search results depends on the user's subjective assessment. As a result, the empirical analysis of a keyword retrieval system's effectiveness is important. during this paper, we tend to examine the evolving practices and resources for effectiveness analysis of keyword searches on relative databases. We compare practices with the longer-standing full-text analysis methodologies in data retrieval. within the

@IJAERD-2014, All rights Reserved

lightweight of this comparison, we tend to build some suggestions for the long run development of the art in evaluating keyword search effectiveness.

Keyword search on unstructured text information has long been studied within the data retrieval community, wherever it goes underneath the name of free text search. Keyword searches offer solely associate degree approximate specification of the knowledge things to be retrieved. Therefore, the correctness of the retrieval can't be formally verified, because it will with question languages like SQL. Instead, retrieval effectiveness is measured by user perception and skill. The empirical assessment of keyword-based retrieval systems is so imperative. The field of keyword search on structured information is well poised for growth towards maturity. the elemental technical and formal issues of performing arts such search are solved , and plenty of vital theoretical results are achieved (in, for example, graph theory). Concern is currently turning to queries of the end-user effectiveness of such search systems. Ancient IR similarity metrics are ported to the new domain, and combined with domain-specific structural options. There's conjointly proof of great enhancements in effectiveness, each through developing new ways and standardization existing ones.

V. EFFECTIVE MEASURES

As declared by Fang Liu, Clement Yu, Weiyi Meng, Abdur Chowdhury, In IR, there are many measures to evaluate effectiveness. 11- point precision and recall (precision is the number of relevant documents retrieved divided by the number of retrieved documents, and recall is the number of relevant documents retrieved divided by the number of relevant documents) is a standard measure. At each of the 11 recall levels (0, 0.1, 0.2…1), a precision value is computed. These eleven exactness are sometimes planned in a very graph for example the effectiveness similarly because the exchange between precision and recall. Mean average exactness (MAP) is another normal live. A exactness is computed when every relevant document is retrieved. Then we tend to average all exactness values to urge one variety to live the effectiveness. Existing IR ways are inadequate in ranking relative outputs. During this paper, we tend to propose a unique IR ranking strategy for effective keyword search. we tend to are the primary that conducts comprehensive experiments on search effectiveness employing a world info and a collection of keyword queries collected by a significant search company. Keyword search permits non-expert users to search out text data in relative databases with far more flexibilities. we tend to planned a unique ranking strategy for effective keyword search in relative databases[5].

VI. CONCLUSION

Keyword search on structured information is so at roughly an equivalent stage that data retrieval was pre- TREC; or, to be less sanguine, the stage that data retrieval had reached by the time of life, and wasn't to obviously surpass for an additional 20 years. there's a lot of promise within the field, however a lot of has to be done to line it on a firm basis, to validate its results, and to inspire the arrogance required to convert this analysis technology into deployed tools. And most of those desiderata rely on enhancements in analysis methodology. the sector of keyword search could, however, still be too young, and therefore the technology too fluid, for a full TREC- vogue cooperative experiment to be doable or maybe acceptable. Instead, the approach forward would appear to be for individual analysis teams to form a lot of thorough, believably freelance, and re-usable check collections, incorporating all 3 elements – corpus, topics. Such associate degree endeavor needs a non-trivial quantity of effort. Keyword search permits non-expert users to seek out text data in relative databases with rather more flexibilities we tend to planned a completely unique ranking strategy for effective keyword search in relative databases. The system generates all answers (tuple trees) for the question. The system computes a ranking score for every answer and ranks them. [3] Finally, top answers area unit came with linguistics. Our ranking strategy is novel. It identifies and uses four new normalization factors that area unit vital to look effectiveness.

The user must enter the appropriate keyword to access the specific content of the database. Again to access the particular content from the whole database user needs to give input through the textbox provided. This is nothing but the query. A query in our model is solely an inventory of keywords, and doesn't got to specify any relation or attribute names. the solution to such a question consists of a rank of "tuple trees," that probably embrace tuples from multiple relations that square measure combined via joins. To rank tuple trees, we have a tendency to introduced a ranking perform that leverages and extends the power of recent electronic database systems to produce keyword search on individual text attributes and rank tuples consequently. With the expansion of the net, there has been a fast increase within the range of users World Health Organization got to access on-line databases while not having close information of the schema or of question languages; even comparatively easy question languages designed for non-experts square measure too difficult for them. We have a tendency to describe BANKS, a system that permits keyword-based search on relative databases, at the side of information and schema browsing.

@IJAERD-2014, All rights Reserved

REFERENCES

[1] Xiaogang Wang, Ke Liu2, Xiaoou Tang2, "Web Image Re-Ranking Using Query-Specific Semantic Signatures", IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume:36 , Issue: 4) April 2014, DOI:10.1109/TPAMI.201

[2] A. Baid, I. Rae, J. Li, A. Doan, and J. Naughton, "Toward Scalable Keyword Search over Relational Data," Proceedings of the VLDB Endowment, vol. 3, no. 1, pp. 140–149, 2010

[3] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan, "Keyword Searching and Browsing in Databases using BANKS," in Proceedings of the 18th International Conference on Data Engineering, ser. ICDE '02, February 2002, pp. 431–440.

[4] Vagelis Hristidis, Luis Gravano, Yannis Papakonstantinou, "Efficient IR-Style Keyword Search over Relational Databases", Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003

[5] Fang Liu, Clement Yu, Weiyi Meng, Abdur Chowdhury, "Effective Keyword Search in Relational Databases", SIGMOD 2006, June 27-29, 2006, Chicago, Illinois, USA

[6] Guoliang Li, Beng Chin Ooi2, Jianhua Feng1,Jianyong Wang, Lizhu Zhou," EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data", *SIGMOD'08,* June 9–12, 2008, Vancouver, BC, Canada.

[7] William Webber," Evaluating the Effectiveness of Keyword Search", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.

@IJAERD-2014, All rights Reserved