National Conference on Sustainable Engineering And Management (NCSEM24) ISSN 2349-6002

Robust Sensitive Information BasedScientific Data Search Engine

Monisha Prabhakaran. P[1]

[1] *Assistant Professor, Computer Science and Engineering, Sivaji College of Engineering and Technology Manivila.*

Abstract- Propose an efficient, secure and fast data searching scheme which will also capable of handling efficient data management. The proposed scheme is used to enable fast search while ensuring the security. Ranking of the document for displaying the document will improves the efficiency. The efficiency of searching also improved by the proposed search engine based scheme. The security can be enhanced by encrypting each of the documents with different keys. This technique keeps the related documents in the same domain so that searching of documents becomes more efficient in terms of time complexity and cost sensitive.

I. INTRODUCTION

Search engine (SEs) are arguably the largest data management systems in the world; although there are larger databases in total storage there is nothing close in query volume. A model search engine handles over 3 billion documents, involving on the order of 10 TB of data, and handles upwards of 150 million queries per day, with peak of several thousand queries per second. This retrospective is based primarily on almost 9 years of work on the Inktomi search engine, from the summer of 1994 through the spring of 2003. It also reflects some of the general issues and approaches of other major search engine – in particular, those of Alta Vista, Info seek and Google – although their actual specific might differ greatly from the examples here. Although queries tend to be short, there are more than ten million different words in nearly all languages. This is a challenge for two reasons. First, implies tracking and ranking ten million distinct words in three billion documents including the position and relative importance (e. g title words) of every word. Second, with few words per query, most queries returns thousands of hits and ranking these hits becomes the primary challenge.

Finally, search engines must be highly available and fresh, two complex and challenging data management issues. Downtime contributes directly to lost revenue and customer churn. Freshness is the challenge of keeping the index up to date with the data, nearly all of which is remote and relatively awkward to access. Most data is "crawled" using the HTTP protocol and some automation, although there is also some data exchange via XML. Despite the size and complexity of these systems, they make almost no use of DBMS systems.

Most of the search engines search for keywords to answer the queries from users. The search engines usually search web pages for the required information. However they filter the pages from searching unnecessary pages by using advanced algorithms. These search engines can answer topic wise queries efficiently and effectively by developing state-of art algorithms. However they are vulnerable in answering intelligent queries from the user due to the dependence of their results on information available in web pages. The main focus of these search engines is solving these queries with closeto accurate results in small time using much researched algorithms. However, it shows that such search engines are vulnerable in answering intelligent queries using this approach. They either show inaccurate results with this approach or show accurate but (could be) unreliable results. With the keywords based searches they usually provide results from blogs (if available) or other discussion boards. The user cannot have a satisfaction with these results due to lack of trusts on blogs etc. To overcome this problem in search engines to retrieve relevant and meaningful information intelligently, sematic web technology deals with a great role. Intelligent semantic technology gives the nearer to desired results by search engines to the user.

One important goal of the semantic web is to make the meaning of information explicitthrough semantic mark-up, thus enabling more effective access to knowledge contained in heterogeneous information environments, such as the web. Semantic search plays an important role inrealizing this goal, as it promises to produce preciseanswers to user's queries by taking advantage of the availability of explicit semantics of

information. For example, when searching for news stories about ph.D students, with traditional searching technologies, I often could only get news entries in which the term "ph.D students" appears. Those entries which mention the names of students but do not use the term "ph.D students" directly will be missed out. Such news entries however are often the once that the user is really interested in. Inthe context of the semantic web, where themeaning of web content is made explicit, the semantic meaning of the keyword (which is a general concept in the example of ph.D students) can be figured out. Furthermore, the underlying semantic relations of metadata can be exploited to support the retrieving of information which is closely related to the keyword. Thereby, the search performance can be significantly improved by expanding the query with instances and relations.

II. LITERATURE SURVEY

A Synonym Based Approach of Data Mining in Search Engine Optimization presented a search based on synonyms. This approach can be further extended or improved by implementing synonym table in a more effective way which should include less space consumption and minimum access time. This SEO approach can increase the ranking of website on SERP and benefit its owners and provide users with more accurate and relevant search results. Advantage is provides better results. Disadvantage is high time consumption.

SemSearch: A Search Engine for the Semantic Web presents Sem Search, a search engine, which pays special attention to this issue by hiding the complexity of semantic search from end users and making it easy to use and effective. In contrast with existing semantic-based keyword search engines which typically compromise their capability of handling complex user queries in order to overcome the problem of knowledge overhead, Sem Search not only overcomes the problem of knowledge overhead but also supports complex queries. Advantage is performance is good and disadvantage is not accurate.

Combining Systems and Database: A Search Engine Retrospective present how a search engine should have been designed in hindsight. Although much of the material has not been presented before, the contribution is not in the specific design, but rather in the combination of principles from the largely independent disciplines of "systems" and "databases". Thus we present the design using the ideas and vocabulary of the database community as a model of how to design data-intensive systems. We then draw some conclusions about the application of data-base principles to other "out of the box" data-intensive systems. Advantage is it is easy to compute and its disadvantages is DBMS for a search engine is that there is a semantic mismatch.

Optimizing Search Engines using Click through Data presents an approach to automatically optimizing the retrieval of search engines using click through data. Intuitively, a good information retrieval system should present relevant documents high in the ranking, with less relevant documents following below. While previous approaches to learning retrieval functions from examples exist, they typically require training data generated from relevance judgments by experts. This makes them difficult and expensive to apply. The goal of this paper is to develop a method that utilizes click through data for training. Advantage is The basic search engines provide a basis for comparison. Disadvantage is information retrieval, rankings need to be consistent only within a query, but not between queries.

Recovering Semantics of Tables on the Web describes a system that attempts to recover the semantics of tables by enriching the table with additional annotations. Their annotations facilitate operations such as searching for tables and finding related tables. Advantage is Performance is good and its disadvantage is not accurate

III. SYSTEM OVERVIEW

I propose an efficient, secure and fast data searching scheme which will also capable of handling efficient data management. The proposed scheme is used to enable fast search while ensuring the security. Ranking of the document for displaying the document will improves the efficiency. The efficiency of searching also improved by the proposed search engine based scheme. The security can be enhanced by encrypting each of the documents with different keys.

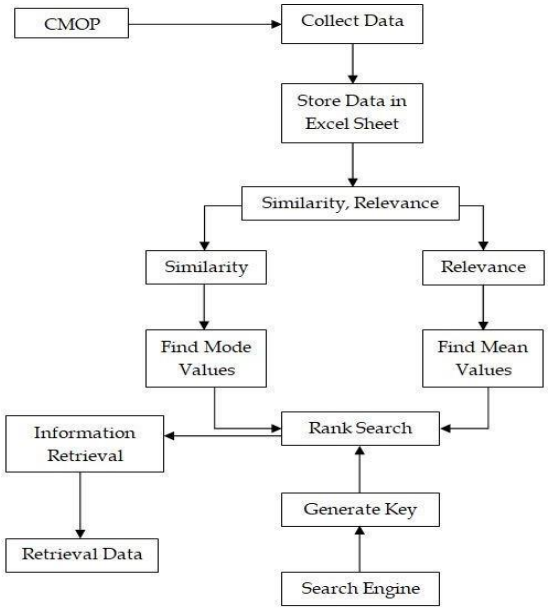
IV. LANGUAGE STUDY - .NET

The Microsoft .NET Framework is a software framework that can be installed on computers running Microsoft Windows operating system. It includes a large library of coded solutions to common programming problems and a virtual machine that manages the execution of programs written specifically for the framework. The .NET Framework is a Microsoft offering and is

intended to be used by most new applications created for the Windows platform.

The .NET Framework family also includes two versions for mobile or embedded device use. A reduced version of the Framework, is available on Windows CE platforms, including Windows Mobile devices such as smartphones. Additionally, the .NET Framework is targeted at severely resource constrained devices.

V. ARCHITECTURE DIAGRAM



VI. SYSTEM IMPLENTATION

Each module has its own importance and they have different tasks, those are listed below:

A. Data collection

The data collection module collects data from Centre for Coastal Margin Observation and Prediction data near here data.

B. Similarity and relevance

In the step first find mode and mean values for similarity calculation. We calculate a similarity score by calculating the distance from that center to the closest distance and the furthest distance of the footprint, averaging them, and scaling them by the size of the range. There is general agreement about what is considered “closer” between collections of data, at least with respect to time and space, and we can approximate this distance (or similarity) in a simple way. We used a four-point scale (3 high relevance, 0 no relevance) relevance is the applicability of the data set’s contents to the searcher’s search.

C. Search engine

In this step first provide key for each document. Searches documents and files for keywords and returns the results of any files containing those keywords. Search engines search for keywords to answer the queries from users.

D. Ranked search

Ranked search of scientific data sets via a relatively straight forward similarity measure. Applying IR techniques to data set ranked search.

E. Information retrieval

IR techniques to scientific-data set search, we need three things: a way to express a scientific information need as a set of search conditions; a method for extracting features from data sets; and a similarity measure to compare search conditions to the extracted features.

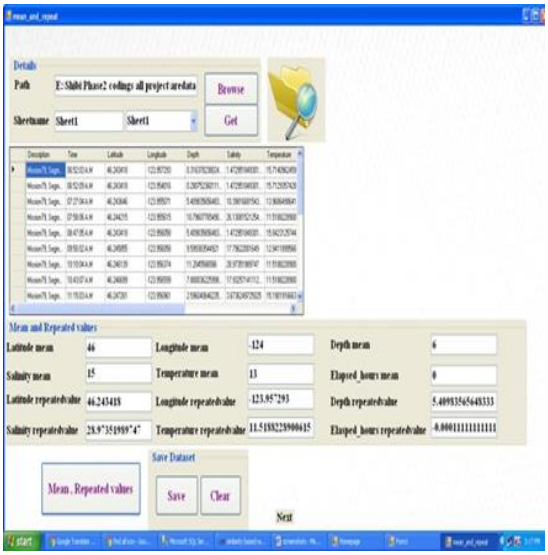
F. Performance analysis

In performance analysis of the proposed system is compared with existing system. In proposed system used search engine so improve the security.

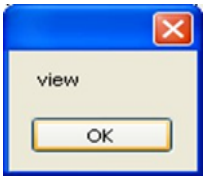
VII. RESULT

A. DATA COLLECTION

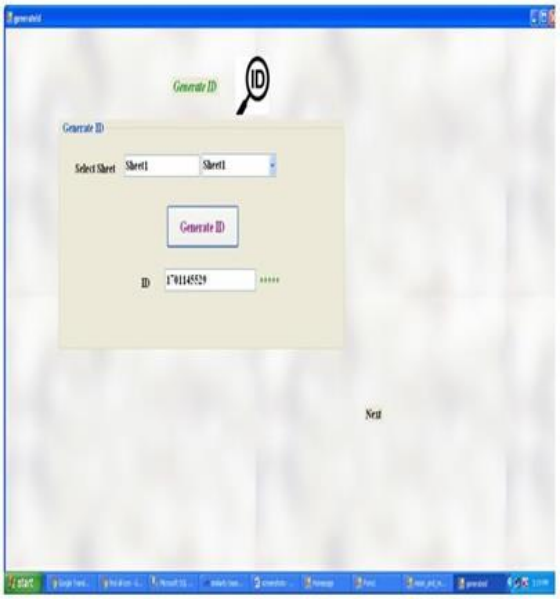
B. MEAN AND REPEATED VALUES



C. VIEW DATA & INSERT DATA

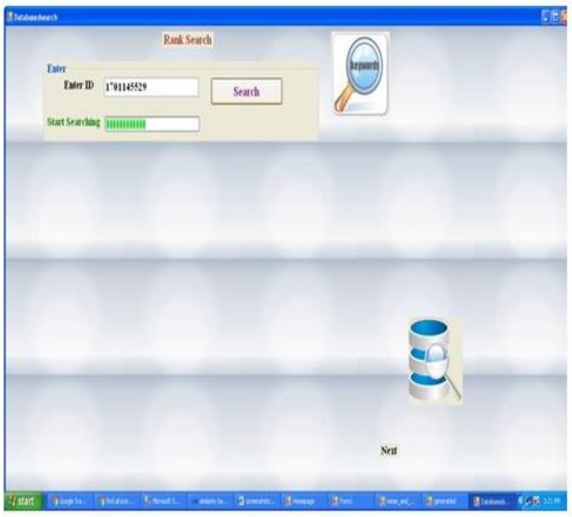


D. GENERATE ID

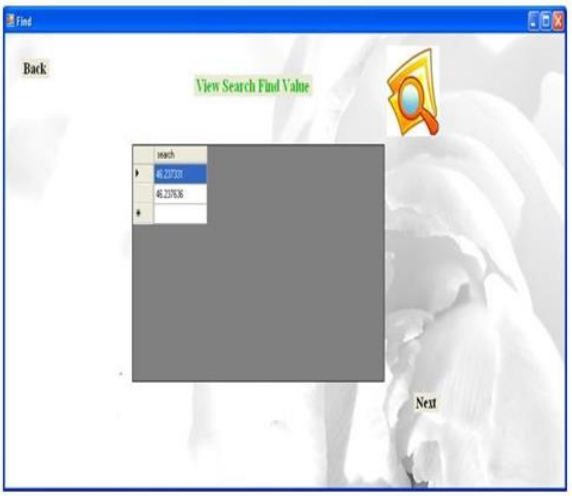


E. ID CREATE

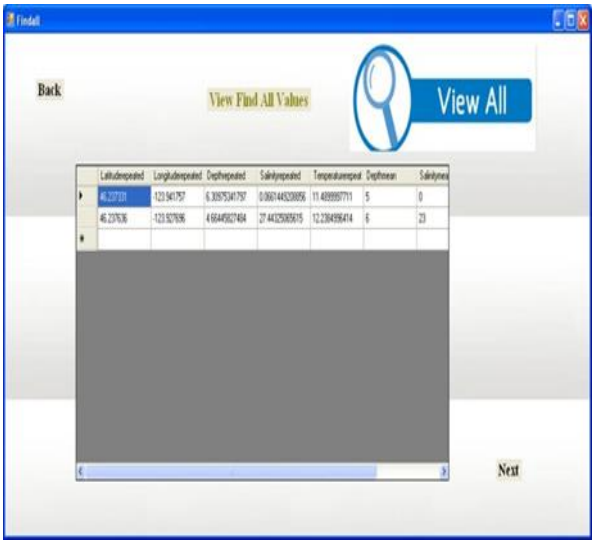
F. RANK SEARCH



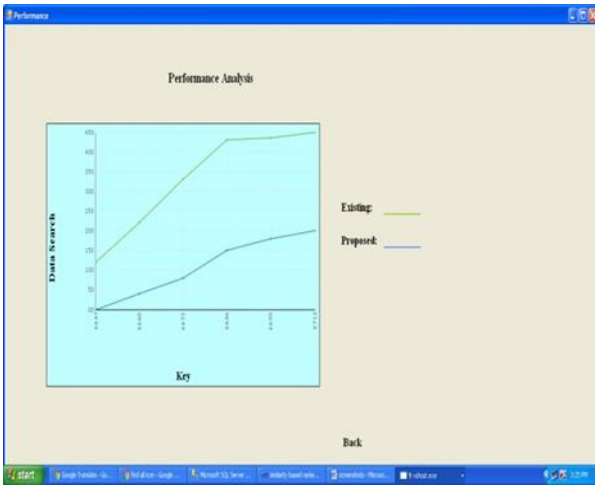
G. FIND



H. FIND ALL VALUES



I. PERFORMANCE ANALYSIS



VIII. CONCLUSION AND FUTURE WORK

This paper presented an approach to Sensitive data based search engine with the goal of improving their retrieval performance automatically. The key insight is that such click through data can provide training data in the form of relative preferences. Based on a new formulationof the learning problem in information retrieval, this paper derives an algorithm for learning a ranking function. Taking a Support Vector approach, the resulting training problem is tractableeven fir large number of queries and large numbers of features. Experimental results show that the algorithm performs well in practice, successfully adapting the retrieval function of a meta-search engine to the preferences of a group of users.

IX. REFERENCES

1.P.Lord and A . Macdonald , "e-Science curation report : Data curation fo e-Science in the UK: An audit to establish requirements for further curation and provition" http://www.jise.ac.uk/uploa-ded_documents /e-ScienceReportFinal.pdf,2003.

2.S.Weidman and T.Arrison , Steps towards Large – Scale Data Integra-tion in the sciences: Summary of Work Shop. Washington, DC, USA: Nat.Acad.press,Aug.2009.

3.J.K.Batcheller, "Automating Geospatial metadata generation ," Comput.geosci.,vol.34,no.4,pp.387-398,2008.

4.A.D'Ulizia, F. Ferri, A. Formica, and P. Grifoni, "Approximating geographical queries," J.Comput. Sci. Technol., Vol. 24,no.6,pp.1109-1124,2009.

5.T.Saracevic, "Relevance : A review of the literature and frame-for thinking on the notion in information science. Parts II,III," J.Amer. Soc.Inform.Sci.Technol.,vol.58,no. 13,pp.2126-2144,2007.