International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.1, Issue.2, pp-347-351 ISSN: 2249-6645

Efficient Query Optimizing System for Searching Using Data Mining Technique

Velmurugan.N
Assistant Professor, Department of MCA,
Saveetha Engineering College,
Thandalam,Chennai-602 105.

Vijayaraj.A
Associate Professor, Department of IT
Saveetha Engineering College,
Thandalam,Chennai-602 105.

ABSTRACT

There is a critical need to design and develop tools that abstract away the fundamental complexity of XML based Web services specifications and toolkits, and provide an elegant, intuitive, simple, and powerful query based invocation system to end users. Web services based tools and standards have been designed to facilitate seamless integration and development for application developers. As a result, current implementations require the end user to have intimate knowledge of Web services and related toolkits, and users often play an informed role in the overall Web services execution process .We employ a set of algorithms and optimizations to match user queries with corresponding operations in Web services, invoke the operations with the correct set of parameters, and present the results to the end user. Our system uses the Semantic Web and Ontologies in the process of automating Web services invocation and execution. Every user has a distinct background and a specific goal when searching for information on the Web. The goal of Web search personalization is to tailor search results to a particular user based on that user's interests and preferences. Effective personalization of information access involves two important challenges: accurately identifying the user context and organizing the information in such a way that matches the particular context. We present an approach to personalized search that involves building models of user context as ontological profiles by assigning implicitly derived interest scores to existing concepts in domain ontology.

Keywords: intuitive, seamless, optimizations, semantic Web, ontology, tailor search.

Introduction

A spreading activation algorithm is used to maintain the interest scores based on the user's ongoing behavior. Our experiments show that re-ranking the search results based on the interest scores and the semantic evidence in an ontological user profile is effective in presenting the most relevant results to the user. With the tremendous growth of information available to end users through the Web, search engines come to play ever a more critical role. Nevertheless, because of their general purpose approach, it is always less uncommon that obtained result sets provide a burden of useless pages. Next generation Web architecture, represented by Semantic Web, provides the layered architecture possibly allowing to overcome this limitation. Several search engines have been proposed, which allow to increase information retrieval accuracy by exploiting a key content of Semantic Web resources, that is relations. However, in order to rank results, most of the existing solutions need to work on the whole annotated knowledge base.

we propose a relation-based page rank algorithm to be used in conjunction with Semantic Web search engines that simply relies on information which could be extracted from user query and annotated resource. Relevance is measured as the probability that retrieved resource actually contains those relations whose existence was assumed by the user at the time of query definition. We address the problem of supporting efficient yet privacy-preserving fuzzy keyword search services over encrypted cloud data. Specifically, we have the following goals: i) to explore new mechanism for constructing storage efficient exact keyword sets; ii) to design efficient and effective fuzzy search scheme based on the constructed keyword sets; iii) to validate the security of the proposed scheme.

Existing System

Searches for the web pages of a person with a given name constitute a notable fraction of queries to Web search engines. A query would normally return web pages related to several namesakes, who happened to have the queried name, leaving the burden of disambiguating and collecting pages relevant to a particular word (from among the namesakes) on the user. Many dynamically generated sites are not indexable by search engines; this phenomenon is known as the invisible web. Some search engines do not order the results by relevance, but rather according to how much money the sites have paid them. Some sites use tricks to manipulate the search engine to display them as the first result returned for some keywords. This can lead to some

search results being polluted, with more relevant links being pushed down in the result list.

Proposed System

We develop web Search approach that clusters web pages based on their association to different people. Our method exploits a variety of semantic information extracted from web pages, such as named entities and hyperlinks, to disambiguate among namesakes referred to on the web pages. We demonstrate the effectiveness of our approach by testing the efficiency of the disambiguation algorithms and its impact on person search. Our system uses word variants or stemming technology, which not only searches for the words present in the user query but also for similar words. This is implemented by domain independent technologies like thesaurus matching as well as by the use of Semantic Web and ontology technologies.

Feasibility study:

A feasibility study is an evaluation of a proposal designed to determine the difficulty in carrying out a designated task. Generally, a feasibility study precedes technical development and project implementation.

Technology and system feasibility:

The assessment is based on an outline design of system requirements in terms of Input, Processes, Output, Fields, Programs, and Procedures. This can be quantified in terms of volumes of data, trends, frequency of updating, etc. in order to estimate whether the new system will perform adequately or not. This means that feasibility is the study of the based in outline.

Economic feasibility:

Economic analysis is the most frequently used method for evaluating the effectiveness of a new system. More commonly known as cost/benefit analysis the procedure is to determine the benefits and savings that are expected from a candidate system and compare them with costs. If benefits outweigh costs, then the decision is made to design and implement the system. An entrepreneur must accurately weigh the cost versus benefits before taking an action. Time Based: Contrast to the manual system management can generate any report just by single click .

Cost Based: No special investment is needed to manage the tool. No specific training is required for employees to use the tool. Investment requires only once at the time of installation. The software used in this project is freeware so the cost of developing the tool is minimal

Legal feasibility:

Determines whether the proposed system conflicts with legal requirements, e.g. a data processing system must comply with the local Data Protection Acts.

Operational feasibility:

Is a measure of how well a proposed system solves the problems, and takes advantages of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development.

Schedule feasibility:

A project will fail if it takes too long to be completed before it is useful. Typically this means estimating how long the system will take to develop, and if it can be completed in a given time period using some methods like payback period. Schedule feasibility is a measure of how reasonable the project timetable is. Given our technical expertise, are the project deadlines reasonable? Some projects are initiated with specific deadlines. You need to determine whether the deadlines are mandatory or desirable.

Market and real estate feasibility:

Market Feasibility Study typically involves testing geographic locations for a real estate development project, and usually involves parcels of real estate land. Developers often conduct market studies to determine the best location within a jurisdiction, and to test alternative land uses for a given parcels. Jurisdictions often require developers to complete feasibility studies before they will approve a permit application for retail, commercial, industrial, manufacturing, housing, office or mixed-use project. Market Feasibility takes into account the importance of the business in the selected area.

Resource feasibility:

This involves questions such as how much time is available to build the new system, when it can be built, whether it interferes with normal business operations, type and amount of resources required, dependencies, etc. Contingency and mitigation plans should also be stated here.

SYSTEM DESIGN

OVERVIEW OF DESIGN

Design is multi-step process that focuses on data structure software architecture, procedural details, and interface between modules. Design is the place where quality is fostered in software engineering. Design is the perfect way to accurately translate a customer's requirement in to a finished software product. The design of an information system produces the details that state how a system will meet the requirements identified during analysis. The emphasis is on translating the performance, requirements into design specifications. The various steps

involved in designing the **"Step Construction Using Visual Cryptography Schemes"** are given below.

- First, decide how the output is to be produced in what format.
- Second, the input data can communicate with applications have to be designed based on the requirements.
- Finally, details related to the justification of the system to be presented.

INPUT DESIGN

It is the process of converting input data to the computer-based data. The goal of designing is to make data entry as easier and free from error as possible. Input design determines the format and validation criteria for data entering the system. Personal computers and terminals can place a data at user's finger tips, allowing them to call up specific data and make timely decisions based on the data.

This system contains data collection screen which display heading the defined their purpose. By employing flashing error messages, and providing necessary alerts on the screen, mist entering of data in the system is avoided.

OUTPUT DESIGN

Computer output is the most important and the direct source of information to the user. Efficient and intelligible output design should improve the system relationships with the user and help in decision making.

Major forms and Web Pages of output are hard copy from the printer and the soft copy from the CRT Display. Output is the key tool to evaluate the performance of software so the designing of output should be done with great care. It should be able to satisfy the user's requirements.

CODE DESIGN

A group of characters used to identify and item of data is a code. A major problem encounter in working with a large amount of data is the retrieval of specific dada when it is required. Code facilitated easier identification simplification in handling and retrieval of item by In the developed system a suitable coding is adopted, which can identify each user exactly.

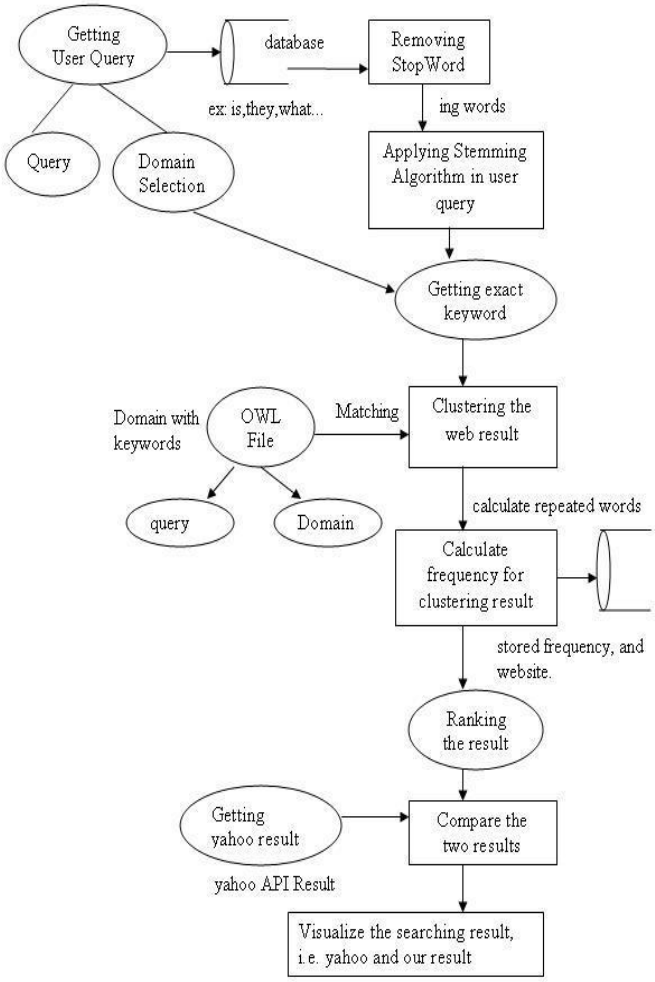

Fig: Data Flow Diagram

Module Description:

- Getting user i/p and stem the keyword
- Web page information retrieval
- Clustering and ranking the web pages
- Implement user profile searching
- Implement relation based searching
- Compare search result

Getting user i/p and stem the keyword

A user submits a query to the middleware via a specialized Web-based interface. User input is involved in various process such as stop word remover, stemming etc. The user query divided in to stopword and addword. Stop word is nothing but which,they,what,where etc… addword is nothing but keywords. The search engine got that keywords and process is performed. Finally we apply the steeming algorithm and get the stem word.This Algorithm attempt to reduce a word to its stem or root form. Thus, the

key terms of a query or document are represented by stems rather than by the original words. This not only means that different variants of a term can be conflated to a single representative form – it also reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time. the words that appear in documents and in queries often have many morphological variants. Thus, pairs of terms such as "computing" and "computation" will not be recognised as equivalent without some form of natural language processing (NLP).

Web page information retrieval

The middleware queries a search engine with this query via the search engine API and retrieves a fixed number (top K) of relevant web pages. The retrieved web pages are preprocessed: . TF/IDF. Preprocessing steps for computing TF/ IDF are carried out. They include stemming, stop word removal, noun phrase identification, inverted index computations, etc. Named entities (NEs) and Web related information is extracted from the web pages. Information Retrieval (IR) is essentially a matter of deciding which documents in a collection should be retrieved to satisfy a user's need for information. The user's information need is represented by a query or profile, and contains one or more search terms, plus perhaps some additional information such importance weights. Hence, the retrieval decision is made by comparing the terms of the query with the index terms (important words or phrases) appearing in the document itself. The decision may be binary (retrieve/reject), or it may involve estimating the degree of relevance that the document has to the query.

Clustering and Ranking the pages:

The clustering algorithm takes the graph, TF/IDF values, and model parameters and disambiguates the set of web pages . The result is a set of clusters of these pages with the aim being to cluster web pages based on association to real person. A set of keywords that represent the web pages within a cluster is computed for each cluster. The goal is that the user should be able to find the person of interest by looking at the sketch. All clusters are ranked by a chosen criterion to be presented in a certain order to the user. Once the user hones in on a particular cluster, the web pages in this cluster are presented in a certain order, computed on this step.

Implement user profile searching:

Personalized web search system, which can learn a user's preference implicitly and then generate the user profile automatically. When the user inputs query keywords, more personalized expansion words are generated by the proposed algorithm, and then these words together with the query keywords are submitted to a popular search engine such as Baidu or Google. These expansion words can help search engines retrieval information for a user according to his/her implicit search intentions, and return different search results to different users who input the same keywords.

Implement relation based searching

With the tremendous growth of information available to end users through the Web, search engines come to play ever a more critical role. Nevertheless, because of their general purpose approach, it is always less uncommon that obtained result sets provide a burden of useless pages. Next generation Web architecture, represented by Semantic Web, provides the layered architecture possibly allowing to overcome this limitation. Several search engines have been proposed, which allow to increase information retrieval accuracy by exploiting a key content of Semantic Web resources, that is relations. However, in order to rank results, most of the existing solutions need to work on the whole annotated knowledge base. In this paper we propose a relation-based page rank algorithm to be used in conjunction with Semantic Web search engines that simply relies on information which could be extracted from user query and annotated resource. Relevance is measured as the probability that retrieved resource actually contains those relations whose existence was assumed by the user at the time of query definition.

Compare search result

Compare existing yahoo result for users given query and our modern relation **s**earch result for users given query and property matching and visualize the results in both search engine based on indexing.

Conclusion

We formalize and solve the problem of supporting efficient yet privacy-preserving fuzzy search for achieving effective utilization of remotely stored encrypted data in Cloud Computing. We design an advanced technique (i.e., wildcard-based technique) to construct the storage-efficient fuzzy keyword sets by exploiting a significant observation on the similarity metric of edit distance.

Future Enhancement

Based on the constructed fuzzy keyword sets, we further propose an efficient fuzzy keyword search scheme. Through rigorous security analysis, we show that our proposed solution is secure and privacy-preserving, while correctly realizing the goal of fuzzy keyword search. we will continue to research on security mechanisms that support: 1) search semantics that takes into consideration conjunction of keywords, sequence of keywords, and even the complex natural language semantics to produce highly relevant search results; and 2) search ranking that sorts the searching results according to the relevance criteria.

REFERENCES

- Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang, Member, IEEE, and Ming-Syan Chen, Fellow, IEEE
- C.C. Aggarwal, A. Hinneburg, and D. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," Proc. Eighth Int'l Conf. Database Theory (ICDT), 2001.
- C.C. Aggarwal and C. Procopiuc, "Fast Algorithms for Projected Clustering," Proc. ACM SIGMOD Int'l Conf. Management of Data, 1999.
- C.C. Aggarwal and P.S. Yu, "The IGrid Index: Reversing the Dimensionality Curse for Similarity Indexing in High Dimensional Space," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2000.
- [R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," Proc. ACM SIGMOD Int'l Conf.
- Management of Data, 1998.

I. Assent, R. Krieger, E. Muller, and T. Seidl, "DUSC: Dimensionality Unbiased Subspace Clustering," Proc. IEEE Int'l Conf. Data Mining (ICDM), 2007.

- K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is Nearest Neighbors Meaningful?" Proc. Seventh Int'l Conf. Database Theory (ICDT), 1999.
- Blum and P. Langley, "Selection of Relevant Features and Examples in Machine Learning," Artificial Intelligence, vol. 97, pp. 245-271, 1997.
- M.-S. Chen, J. Han, and P.S. Yu, "Data Mining: An Overview from Database Perspective," IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, pp. 866-883, Dec. 1996.
- C.H. Cheng, A.W. Fu, and Y. Zhang, "Entropy-Based Subspace Clustering for Mining Numerical Data," Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 1999.
- www.w3schools.com

Authors Profile:

A.Vijayaraj is an Associate Professor in Department of Information Technology at Saveetha Engineering College. He received his Master of Computer Application in Bharathidhasan University, in 1997 and his Master of Engineering in Computer Science and Engineering from Sathyabama University at 2005. He has 12 years of teaching experience from various Engineering Colleges during tenure he was Awarded **Best Teacher Award** twice..He is a Member of, CSI and ISTE. He has Published 2 papers in International journal and 10 Papers in International and National Level conferences. His area of interest includes Operating Systems, Data Structures, Networks and Communication

N.Velmurugan is an Asst.Professor in Department of Computer Applications at Saveetha Engineering College. He received his Master of Computer Application in Bharathidhasan University, in 1999 and his Master of Engineering in Computer Science and Engineering from Sathyabama University at 2011. He has 11 years of teaching experience from various Engineering Colleges during tenure he was Awarded **Best Teacher Award** .He is a Member of ISTE. He has Published 1 papers in International journal and 5 Papers in International and National Level conferences. His area of interest includes Operating Systems, Data Structures, Network Security and Cryptography.