THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Efficient Search Solution on Unstructured Data

Boraste P. S.
B.E. Computer Student, K.K.W.I.E.E.R. (Nashik), University of Pune, India
Bankar S. D.
B.E. Computer Student, K.K.W.I.E.E.R. (Nashik), University of Pune, India
Junagade A. P.
B.E. Computer Student, K.K.W.I.E.E.R. (Nashik), University of Pune, India
Bahalkar G. S.
B.E. Computer Student, K.K.W.I.E.E.R. (Nashik), University of Pune, India
Sharma N. G.
Assistant Professor, K.K.W.I.E.E.R., India

Abstract:
Markup languages are used to store and share data, but due to the undefined structure, searching may lead to ambiguous results. Keyword searching is generally used for retrieving the relevant data from large web database. Web databases may be in any one of the forms XML, JSON, AdsML, AIML, APM L, ATML, BeerXML, SGML etc. XML is a markup language used to represent data structures like records, lists, trees. Redundant data occurs in XML and it leads to an increase in storage along with an increased cost for data transfer and manipulation. Since JSON parses unstructured data much faster as compared to XML and is more compact. The ambiguity that results from searching unstructured data can be avoided by using baseline and anchor pruning algorithms. As a Baseline and anchor pruning algorithms are used on XML to enhance search results and to improve accuracy. These algorithms are applied parallely on XML and JSON to get top-K desired and accurate results. Baseline algorithm uses diversification method to enhance search results. Anchor-pruning algorithm is used to improve accuracy. These results may help user to select relevant queries and modify them as per requirement.

Keywords: *XML, JSON, Keyword search, SLCA, Novelty Ranking, Query disambiguation, Semantic search, context based diversification.*

Problem Definition: *To design context based diversification model for the given keyword query over unstructured data to obtain multiple search pattern and top-K match result.*

1. Introduction

Extensible markup language is used to describe data. XML standard is a flexible way to create information format. It is used for scientific data to annotate large documents. XML format can represent both structured and unstructured data. Keyword searching is also important to query XML data with regular structure. It is generally used for retrieving relevant data from large web database. Redundant data occurs in XML. Redundancy stored information can lead not only higher data storage cost but also to increase cost for data transfer and data manipulation. Compared with keyword search with information retrieval that prefer to find list of relevant documents, keyword search approaches in structured and semi structured data concentrate more on specific information contents. e.g. smallest lowest common ancestor (SLCA) based keyword search in XML. If node is SLCA then its ancestor will be excluded from SCLCAs by which the minimal information content with SLCA semantics can be used to represent the specific results in XML keyword search.

JSON parses unstructured data much faster as compare to XML and is more compact. Markup languages are used to store and share data, but due to the undefined structure, Searching leads to ambiguous results.

In this paper diversification method is proposed, By using this method more unambiguous results may obtained. By using anchor based pruning algorithm, Top-K desired and efficient result is to be obtained.

In general, user can select relevant queries form Top-K result or can modify query according to user's need.

2. Literature Survey

2.1. Results Diversification for XML Keyword Search Based on the Semantic Category of Central Entity

In recent years, result diversification for keyword search has attracted increasing attention. This paper addressed the problem of results diversification based on the semantic categories of central entities described by search results creatively. So user can locate the results. This paper focuses on organizing search results by diversifying rather than getting the most relevant search results. So user can locate the results of interest easily navigate. On the other hand, This method categories results independently of information of keywords. Online processing time would be greatly reduced because a plenty of heavy work have been finished off line inputting keywords. This is an obvious advantage over other methods.

2.2. Automatic Metadata Extraction and Classification of Spreadsheet Documents Based on Layout Similarity

Effective information search is becoming a significant success for business. Metadata is an essential part of modern information system since it has proven to be helpful to the people to find related documents from different repositories. The objective of this paper proposed an innovative method that automatically performs selection of metadata, classification based on the spreadsheets having layout same to that of a given sample spreadsheet whose metadata is previously defined. This paper proposed metadata extraction and classification methods. Search index is generated from the classified metadata enables users to define the meanings of the keywords in the search query.

2.3. Improving Search Quality of an XML Context-Driven Search Engine by Relabeling of Canonical Trees

Both Keyword-based and Context-driven are answered by Context-Driven Search Engine queries using stack-based sort merge algorithm. Context-driven querying is nothing but keyword-based querying along with some structural constraints, which helps to search the element. Due to the use of contexts while searching, more relevant information is found as answer. So it results in performance improvement. This paper has presented XML Context-driven Engine Search. It utilizes techniques which are used in XML Context-driven search. In addition to new techniques that hold the type of XML trees which XCD Search does not handle well. This system experimentally and compared with the original version of XML Context-driven Search method. The results showed remarkable improvements. This paper proposed XML search engine called 'Improved XCD Search Engine', which is an improvement over XCD Search.

2.4. Ontology-based Discovery of Information an Application in Directory Service System of Yellow River Data Resources

At present, there was vast amount of data and application with different databases which is available in distributed environment, how to find them is a crucial task. An ontology driven approach is put forward to overcome semantic heterogeneity which is caused by synonyms as well as homonyms during keyword-based search in catalogues. An ontology plays important role in the development of a dimentionally aware search engine, with regard to providing support for query disambiguate, query expansion, relevance ranking. This paper has given a practical case study to what extent the information based on ontology discovery can solve semantic heterogeneity problems.

This paper introduced Ontology technology into the level of knowledge of Yellow River information and integrated with distributed information resources for query.

2.5. A Layered Interchange Scan Algorithm Based on Semantic Context

Along with the rapid growth of xml database on the Internet, the xml database retrieval research has getting more attention. The searching algorithm which is based on the key words is a research hotspot in the field. This paper has given a context-based layered intersection scan algorithm (CLISA) that are Compared with other documents, such as XML documents have lots of advantages, including large amount of data storage cross platform. This paper presented a context-based layered intersections can algorithm. The experiment has demonstrated that this algorithm has an improvement of accuracy comparing with other LISA algorithms. Even through computation efficiency is little affected.

2.6. Structured Based Clustering Technique of XML Documents. (2013)

In this, the structured data contain by XML documents is group in together, in order to form a tree like structure, in this structure of tree assumption of nodes is made and then the redundancy is reduced from the nodes. Thus a sophisticated tree is formed. Thus when user fine a query he/she is likely to get the specific and precise data from the database as the database is converted into tree form and clustering of similar data.

2.7. Research on XML keyword Query Method Based on Semantic. (2013)

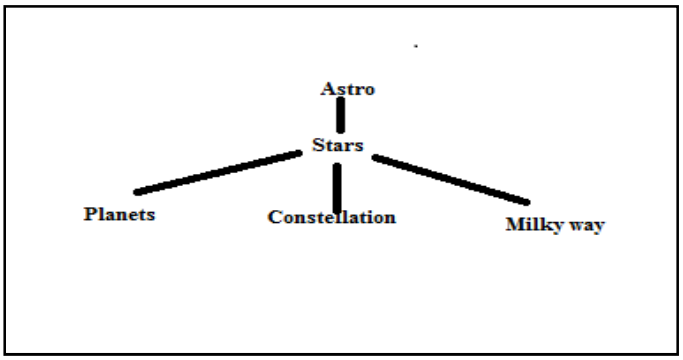

Figure 1

The traditional method to search a keyword is SLCA (Smallest Lowest Common Ancestor) which is carried on structure data. E.g. consider a XML document named as "astro.xml" containing the information of stars, constellation milky way, galaxy.
If a user put galaxy as a keyword then in SLCA if would not get any output. Hence the use of semantic is not made but is SKSA is used the semantic of galaxy that is milky way can be given as output thus the user will be provided with the output as synonym.

2.8. Answering Approximate Queries over XML Data. (2015)
The technique of query relaxing is used the feature of a query is mapped with the structured dataset and retrieval of data contain top-k results which are approximately matching the feature of the query which is fixed by the user and semantic of the words are also kept in considerations.

2.9. An efficient Privacy Conserving Ranked Keyword Search Method
As in our paper we are performing keyword search on XML data. In this paper author chi chen. explains us how to perform key search over cloud data by preserving privacy of data. For that they investigate cipher text search. They explain the problem of manufacturing the semantic relationship which is also there in our XML keyword search. In this paper, they say that: Properly solves the multi-keyword ranked search problems and also improve in search efficiency and rank security.

2.10. Summarizing Search Results with Community Based Question Answering
In Shorten, generation previously main focus is only on how to produce one snippet for individual search result. They generate snippets as a comprehensive overview, e.g. Flu is entity query in a search-result page. The Chung-Lun Chiang, Shih-ring Chen and Pu-Jen Cheng say that their approach is first extracts the attribute for example mark and diagnose which are of categories disease. For community based Question based answering website. In this Integer Linear programming is used to find optimal sentences. This experiment is conducted on Wikipedia and yahoo by calculating experiment results it demonstrates effectiveness of their approach. At the last they also compare it to an existing commercial search engine baseline.

2.11. Top-k Keyword Search over probabilistic XML data (2011)
Due to word Semantic it is (impossible to) search keyword over probabilistic XML data by traditional method. In Top-k keyword search the data is retrieved ask SLCA results with the k highest probability. Then Jain Xin Li, Chegfei Liu, Rui Zhous, wei wang proposed 2 new algorithm 1) Prstack- which can find k SLCA result. 2) At the last they design two algorithms and compare their performance with analysis of experimental result.

2.12. X-clean: Providing Values Spelling Suggesting for XML Keyword Queries (2011)
In this paper author says that for efficiently and effectively search on XML document. The one important approach is to suggest the query to user if user query is having some error. They study the problem of effective and efficient by providing quality suggestions for keyword queries over XML data. In X-clean, different error model are there and due to this keyword query semantic without losing rigor is possible. This algorithm is designed to suggest Top-k suggestions efficiently. They also study two large datasets to prove the effectiveness and efficiency of algorithms. In X-clean, they derived the final scoring function based on the probabilities theory and the state-of-the art language model which is based on – the error model, that model the typing errors, the query generation model that the probability of user being interested in a particular query. XML tree structure and keyword query semantics, which identify until of information in XML data and measure quality of query in finer granularity.

2.13. Diverse Set Selection over Dynamic Data (2014)
Top-k Diverse results are retrieved. Here, the tree data structure is used to show the data it become to provide specific data to user request the insertion and deletion of nodes containing data is a key technique used in proposed algorithm. Due to this there data contain data by the tree changes frequently, but by following insertion and deletion technique the tree becomes more efficient. Thus providing more specific data with Top-k diverse data to the user.

2.14. A Survival Modeling Approach to Biomedical Search Result Diversification Using Wikipedia (2013)

Survival modeling approach to promoting ranking diversity as well as Biomedical Search Result for biomedical information retrieval (IR). This paper is based on the following concepts: get details of the survival modeling for diversification retrieval result which is based on the probabilistic model to get the experimental result, experiment result using Wikipedia (also used for concepts and lexical variants) for aspect detection and filtering. Also Web retrieval environment results the effectiveness of the result, diversification also in Web retrieval in the future.

2.15. Diversifying Results of top-k Queries over Bounded Regions (2014)

Initially to get Top-k diverse query the process required to visit the each and every node if the tree. Which prove to be inefficient data consuming thus to overcome these two techniques are used 1. Euclidean distance space to get minimal distances 2. To get data according to highest priority first in top-k result which provide efficiency to the algorithm? To explore more into make proposed algorithm more efficient batch processing can be used.

3. References

i. Marina Drosou , Evaggelia Pitoura.(2014). Diverse Set Selection over Dynamic Data.
ii. Yu Zhou, Guohua Liu, BiYing Wang.(2014) . Diversifying results of top-k querys over bounded regions.
iii. Xiaoshi Yin, Jimmy Xiangji Huang.(2013). A Survival Modeling Approach to Biomedical Search Result Diversification Using Wikipedia .
iv. Zhikui Chen, Youming Luo , Zhuang Shao.(2010). A Layered Intersection Scan Algorithm Based on Semantic Context.
v. Yan Li, Feixue Li . (2011). Ontology-based Discovery of Information-An Application in Directory Service System of Yellow River Data Resources .
vi. Suhas B. Bhagate.(2011) Improving Search Quality of An XML Context-Driven Search Engine by Relabling of Canonical Trees.
vii. Somchai Chatvichienchai.(2011) . Automatic Metadata Extraction and Classification of Spreadsheet Documents Based on Layout Similarity.
viii. Yuling Song.(2013) .Results Diversification for XML Keyword Search Based on the Semantic Category of Central Entity.
ix. Mary Posonia A.(2013). Structural- based Clustering Technique 0/X ML Documents.
x. Guofeng Zhaoa, Shan Tianb. (2013). Research on XML Keyword Query Method Based on Semantic.
xi. Jian Liu and D.L. Yan.(2015). Answering Approximate Queries over XML Data.
xii. Yifei Lu , Wei Wang , Jianxin Li , Chengfei Liu.(2015). XClean: Providing Valid Spelling Suggestions for XML Keyword Queries.
xiii. Jianxin Li1, Chengfei Liu1, Rui Zhou1, Wei Wang .(2011) .Top-k Keyword Search over Probabilistic XML Data.
xiv. Chung-Lun Chiang, Shih-Ying Chen and Pu-Jen Cheng.(2014). Summarizing Search Results with Community-Based Question Answering.
xv. Chi Chen, Xiaojie Zhu, Peisong Shen.(2014). An Efficient Privacy-Preserving Ranked Keyword Search Method.