International Journal of Applied Sciences, Engineering and Management
ISSN 2320 – 3439, Vol. 04, No. 01, January 2015, pp. 57 – 61

CONTEXT-BASED DIVERSIFICATION FOR KEYWORD QUERIES OVER XML DATA

Sivaji Yerraguntla[1], Borugadda.NagaRaju[2]

[1]M.Tech Student, Dept of CSE, Amrita Sai Institute of Science and Technology, Paritala, Krishna-521180.

[2]Assistant Professor, Dept of CSE, Amrita Sai Institute of Science and Technology, Paritala, Krishna-521180.

Abstract: While keyword query empowers ordinary users to search vast amount of data, the ambiguity of keyword query makes it difficult to effectively answer keyword queries, especially for short and vague keyword queries. To address this challenging problem, in this paper we propose an approach that automatically diversifies XML keyword search based on its different contexts in the XML data. Given a short and vague keyword query and XML data to be searched, we firstly derive keyword search candidates of the query by a simple feature selection model. And then, we design an effective XML keyword search diversification model to measure the quality of each candidate. After that, two efficient algorithms are proposed to incrementally compute top-k qualified query candidates as the diversified search intentions. Two selection criteria are targeted: the k selected query candidates are most relevant to the given query while they have to cover maximal number of distinct results. At last, a comprehensive evaluation on real and synthetic datasets demonstrates the effectiveness of our proposed diversification model and the efficiency of our algorithms

***Index Terms**: XML Keyword Search, Context-based Diversification.*

INTRODUCTION

Keyword search on structured and semi-structured data has attracted much research interest recently, as it enables users to retrieve information without the need to learn sophisticated query languages and database structure. Compared with keyword search methods in Information Retrieval that prefer to find a list of relevant documents, keyword search approaches in structured and semi-structured data (denoted as DB&IR) concentrate more on specific information contents, e.g., fragments rooted at the smallest lowest common ancestor (SLCA) nodes of a given keyword query in XML. Given a keyword query, a node v is regarded as an SLCA if the sub tree rooted at the node v contains all the keywords, and there does not exist a descendant node v′ of v such that the sub tree rooted at v′ contains all the keywords. In other words, if a node is an SLCA, then its ancestors will be definitely excluded from being SLCAs, by which the minimal information content with SLCA semantics can be used to represent the specific results in XML keyword search. In this paper, we adopt the well accepted SLCA semantics as a result metric of keyword query over XML data. In general, the more keywords a user's query contains, the easier the user's search intention with regards to the query can be identified. However, when the given keyword query only contains a small number of vague keywords, it would become a very challenging problem to derive the user's search intention due to the high ambiguity of this type of keyword queries. Although sometimes user involvement is helpful to identify search intentions of keyword queries, a user's interactive process may be time consuming when the size of relevant result set is large. To address this, we will develop a method of providing diverse keyword query suggestions to users based on the context of the given keywords in the data to be searched. By doing this, users may choose their preferred queries or modify their original queries based on the returned diverse query suggestions.

SYSTEM ANALYSIS

EXISTING SYSTEM

Keyword search on structured and semi-structured data has attracted much research interest recently, as it enables users to retrieve information without the need to learn sophisticated query languages and database structure. Compared with keyword search methods in Information Retrieval (IR) that prefer to find a list of relevant documents, keyword search approaches in structured and semi-structured data concentrate more on specific information contents, e.g., fragments rooted at the smallest lowest common ancestor (SLCA) nodes of a given keyword query in XML In general, the more keywords a user's query contains, the easier the user's search intention with regards to the query can be identified. However, when the given keyword query only contains a small number of vague keywords, it would become a very challenging problem to derive the user's search intention due to the high ambiguity of this type of keyword queries. Although sometimes user involvement is helpful to identify search intentions of keyword queries, a user's interactive process may be time consuming when the size of relevant result set is large. To address this, we will develop a method of providing diverse keyword query suggestions to users based on the context of the given keywords in the data to be searched. By doing this, users may choose their preferred queries or modify their original queries based on the returned diverse query suggestions.

PROPOSED SYSTEM

We initiate a formal study of the diversification problem in XML keyword search, which can directly compute the diversified results without retrieving all the relevant candidates. Towards this goal, given a keyword query, we first derive the co-related feature terms for each query keyword from XML data based on mutual information in the probability theory, which has been used as a criterion for feature selection. The selection of our feature terms is not limited to the labels of XML elements.

Each combination of the feature terms and the original query keywords may represent one of diversified contexts (also denoted as specific search intentions). And then, we evaluate each derived search intention by measuring its relevance to the original keyword query and the novelty of its produced results. To efficiently compute diversified keyword search, we propose one baseline algorithm and two improved algorithms based on the observed properties of diversified keyword search results.

The remainder of this paper is organized as follows.

- We introduce a feature selection model and define the problem of diversifying XML keyword search.
- We describe the procedure of extracting the relevant feature terms for a keyword query based on the explored feature selection model.
- We propose three efficient algorithms to identify a set of qualified and diversified keyword query candidates and evaluate them based on our proposed pruning properties.
- We provide extensive experimental results to show the effectiveness of our diversification model and the performance of our proposed algorithms.
- We describe the related work in Section and conclude in Section

PROPOSED SYSTEM ALGORITHMS

The authors proposed efficient algorithms to identify keyword clusters in large collections of blog posts for specific temporal intervals

ANCHOR-BASED PRUNING SOLUTION

Although the anchor-based pruning algorithm can avoid unnecessary computation cost of the baseline algorithm, it can be further improved by exploiting the parallelism of keyword search diversification and reducing the repeated scanning of the same node lists.

DATASET AND QUERIES

We selected some terms based on the following criteria:

- ✓ A selected term should often appear in user-typed keyword queries;
- ✓ A selected term should highlight different semantics when it co-occurs with feature terms in different contexts.

AVERAGE TIME COST OF QUERIES

That in about 3.5 seconds and 2 seconds, respectively. This is because lots of nodes can be skipped by anchor nodes without computation. Another reason is that when the number of suggestions is small, e.g., 5, we can quickly identify the qualified suggestions and safely terminate the evaluation with the guarantee of the upper bound. As such, the qualified suggestions and their diverse results can be output.

BASELINE SOLUTION

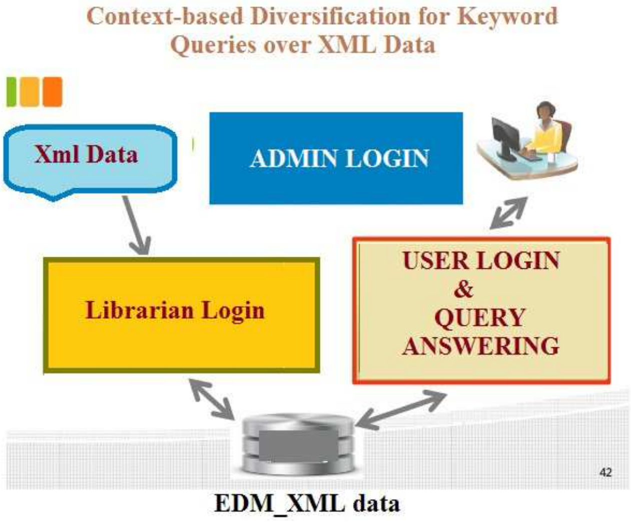
Different from traditional XML keyword search, our work needs to evaluate multiple intended query candidates and generate a whole result set, in which the results should be diversified and distinct from each other. Therefore, we have to detect and remove the duplicated or ancestor SLCA results that have been seen when we obtain new generated results.

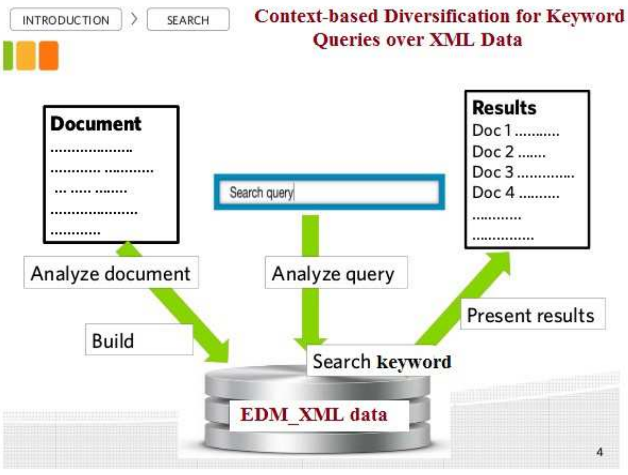
ADVANTAGES

The first work to measure the difference of XML keyword search results by comparing their feature sets. However, the selection of feature set in is limited to metadata in XML and it is also a method of post-process search result analysis. Therefore, the simple measure can be used to quantify how much the observed word co-occurrences maximize the dependency of feature terms while reduce the redundancy of feature terms.

After that, all the generated term pairs will be recorded in the term correlated graph. In the procedure of building correlation graph, we also record the count of each term-pair to be generated from different entity nodes. As such, after the XML data tree is traversed completely, we can compute the mutual information score for each term-pair based on Equation. To reduce the size of correlation graph, the term-pairs with their correlation lower than a threshold can be filtered out. Based on the off-line built graph, we can on-the-fly select the top-m distinct terms as its features for each given query keyword.

SYSTEM ARCHITECTURE


Context-based Diversification for Keyword Queries over XML Data


Context-based Diversification for Keyword Queries over XML Data

RELATED WORK

This work is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things r satisfied, ten next steps is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into account for developing the proposed system.

We have to analysis the Data Engineering.

Data Engineering:

What is Data Engineering?

Data engineering is the multi-disciplinary practice of engineering computing systems, computer software, or extracting information partly through the analysis of data. This note briefly discusses the issues and the disciplines involved and points to some introductory and survey literature.

Remotely hosted:

Services or data are hosted on someone else's infrastructure.

Ubiquitous:

Services or data are available from anywhere.

Co modified:

The result is a utility computing model similar to traditional that of traditional utilities, like gas and electricity. You pay for what you would like.

Software as a Service (SaaS):

SaaS is a model of software deployment where an application is hosted as a service provided to customers across the Internet. SaaS is generally used to refer to business software rather than consumer software, which falls under Web 2.0. By removing the need to install and run an application on a user's own computer it is seen as a way for businesses to get the same benefits as commercial software with smaller cost outlay. Saas also alleviates the burden of software maintenance and support but users relinquish control over software versions and requirements. The other terms that are used in this sphere include *Platform as a Service* (PaaS) and Infrastructure as a Service (IaaS).

Cloud Storage:

Several large Web companies (such as Amazon and Google) are now exploiting the fact that they have data storage capacity which can be hired out to others. This approach, known as 'cloud storage' allows data stored remotely to be temporarily cached on desktop computers, mobile phones or other Internet-linked devices. Amazon's Elastic Compute Cloud (EC^2) and Simple Storage Solution (S3) are well known examples.

- Outsourcing

Despite of the various advantages of cloud services, outsourcing sensitive information (such as e-mails, personal health records, company finance data, government documents, etc.) to remote servers brings privacy concerns. The cloud service providers (CSPs) that keep the data for users may access users' sensitive information without authorization.

These are significant works as it is highly possible that the data owners need to update their data on the cloud server. But few of the dynamic schemes support efficient multi keyword ranked search.

- Data users

Data users are authorized ones to access the documents of data owner. With t query keywords, the authorized user can generate a trapdoor TD according to search control mechanisms to fetch k encrypted documents from cloud server. Then, the data user can decrypt the documents with the shared secret key.

- Cloud server

Cloud server stores the encrypted document collection C and the encrypted searchable tree index I for data owner. Upon receiving the trapdoor TD from the data user, the cloud server executes search over the index tree I, and finally returns the corresponding collection of top- k ranked encrypted documents. Besides, upon receiving the update information from the data owner, the server needs to update the index I and document collection C according to the received information.

Privacy-preserving:

The scheme is designed to prevent the cloud server from learning additional information about the document collection, the index tree, and the query.

1) Index Confidentiality and Query Confidentiality: The underlying plaintext information, including keywords in the index and query, TF values of keywords stored in the index, and IDF values of query keywords, should be protected from cloud server.

2) Trapdoor Unlink ability: The cloud server should not be able to determine whether two encrypted queries (trapdoors) are generated from the same search request.

3) Keyword Privacy: The cloud server could not identify the specific keyword in query, index or document collection by analyzing the statistical information like term frequency. Note that our proposed scheme is not designed to protect access pattern, i.e., the sequence of returned documents.

MODULE DESCRIPTION

ADMIN

The administration includes the performance or management of decision making as well as the efficient organization of people and other resources to direct activities toward common goals and objectives**.** An Administration is an insolvency process which is predominately designed to provide a breathing space from creditor actions. The law and procedure relating to

Administrations and more importantly the procedure into Administration is set out at Schedule.

LIBRARIAN

As information experts, librarians search for and find information, collect and organize information, and implement systems and vehicles that make information easy to access from long or short-range locations. Librarians are trained to find and collect all types of information - books, newspapers, magazines, databases, websites, CDs, videos, government publications and any other type of publicly available data. They are also trained to develop systems to organize and manage this information so that it can be easily retrieved. Librarians design and deliver information services for their client groups as well.

USER

Alternatively referred to as an end-user, a user is any individual who is not involved with supporting or developing a computer or service. A user is another name of an account capable of logging into a computer or service. For example, people who log into the Computer Hope forums are considered a user or member**.**

MODULE DESCRIPTION ADMIN

Course Entry:

Would you like added your course details an entry courses and certificate programs are available to prepare students for careers that require the accurate entry of data, such as those in administrative assisting and word processing. Continue reading to learn more about entry courses and certification programs, as well as get information about potential careers

View:

In a database management system, a view is a way of portraying information in the database. This can be done by arranging the data items in a specific order, by highlighting certain items, or by showing only certain items. For any database, there are a number of possible views that may be specified. Databases with many items tend to have more possible views than databases with few items**.**

Search:

Search engines are programs that search documents for specified keywords and return a list of the documents where the keywords were found. A search engine is really a general class of programs.

Queries – Answering

Usually a computer program, may construct its answers by querying a structured database of knowledge or information, usually a knowledge base. More commonly, we like to search any user document an accessed to display user details and Queries.

MODULE DESCRIPTION LIBRARIAN

Books Entry:

Manually- librarian accessed 'books of original entry' is transferred.

Search:

Search engines are programs that search documents for specified keywords and return a list of the documents where the keywords were found. A search engine is really a general class of programs.

MODULE DESCRIPTION USER

Search:

Search engines are programs that search documents for specified keywords and return a list of the documents where the keywords were found. A search engine is really a general class of programs.

Queries – Answering

Usually a computer program, may construct its answers by querying a structured database of knowledge or information, usually a knowledge base. More commonly, we like to search any user document an accessed to display user details and Queries.

INPUT DESIGN AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:'

- ❖ What data should be given as input?
- ❖ How the data should be arranged or coded?
- ❖ The dialog to guide the operating personnel in providing input.
- ❖ Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user.

Efficient and intelligent output design improves the system's relationship to help user decision-making.

Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

Select methods for presenting information.

Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- ❖ Convey information about past activities, current status or projections of the
- ❖ Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.
- ❖ Confirm an action.

CONCLUSION

In this paper, we first presented an approach to search diversified results of keyword query from XML data based on the contexts of the query keywords in the data. The diversification of the contexts was measured by exploring their relevance to the original query and the novelty of their results. Furthermore, we designed three efficient algorithms based on the observed properties of XML keyword search results. Finally, we verified the effectiveness of our diversification model by analyzing the returned search intentions for the given keyword queries over DBLP dataset based on the nDCG measure and the possibility of diversified query suggestions. Meanwhile, we also demonstrated the efficiency of our proposed algorithms by running substantial number of queries over both DBLP and XMark datasets. From the experimental results, we get that our proposed diversification algorithms can return qualified search intentions and results to users in a short time.

REFERENCES

[1] Y. Chen, W. Wang, Z. Liu, and X. Lin, "Keyword search on structured and semi-structured data,"

[2] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram, "Xrank: Ranked keyword search over xml documents,"

[3] C. Sun, C. Y. Chan, and A. K. Goenka, "Multiway slca-based keyword search in xml data,"

[4] Y. Xu and Y. Papakonstantinou, "Efficient keyword search for smallest lcas in xml databases,"

[5] J. G. Carbonell and J. Goldstein, "The use of mmr, diversitybased reranking for reordering documents and producing summaries",

[6] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results,"

[7] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. B¨uttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation,"

[8] C. O. Sakar and O. Kursun, "A hybrid method for feature selection based on mutual information and canonical correlation analysis,"

[9] N. Bansal, F. Chiang, N. Koudas, and F. W. Tompa, "Seeking stable clusters in the blogosphere,"

[10] Angel and N. Koudas, "Efficient diversity-aware search," in *SIGMOD Conference*, 2011, pp. 781–792.

[11] F. Radlinski and S. T. Dumais, "Improving personalized websearch using result diversification," in *SIGIR*, 2006, pp.691-692

[12] Z. Liu, P. Sun, and Y. Chen, "Structured search result differentiation,"*PVLDB*, vol. 2, no. 1, pp. 313–324, 2009.

[13] E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl, "*DivQ*:diversification for keyword search over structured databases,"in *SIGIR*, 2010, pp. 331–338.

[14] J. Li, C. Liu, R. Zhou, and B. Ning, "Processing xml keyword search by constructing effective structured queries," in *APWeb/WAIM*, 2009, pp. 88–99.

[15] H. Peng, F. Long, and C. H. Q. Ding, "Feature selection based on mutual information: Criteria of max-dependency, maxrelevance, and min-redundancy," *IEEE Trans. Pattern Anal.*

[16] *Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[17] http://www.wsdream.net

[18] www.service-computing.com

[19] www.talkincloud.com

[20] www.sevicecomputing.sys-con.com

[21] www.virtualizationreview.com/Home.aspx

[22] www.thecloudtutorial.com