Information Retrieval by Keyword Query Routing

[1] G.Prasanna [2] B Rasagna, [3] Prasad B
[1]II/IV, [2][3]Associate Professor
[1][2][3] Department of CSE, Marri Laxman Reddy Institute of Technology and Management (MLRITM) Hyderabad
[1] prasannaguduri19@gmail.com [2] bheemarasagna@gmail.com [3]bprasad@gmail.com

Abstract: **Keyword search is an intuitive paradigm for searching linked data sources on the web. We propose to route keywords only to relevant sources to reduce the high cost of processing keyword search queries over all sources. We propose a novel method for computing top-k routing plans based on their potentials to contain results for a given keyword query. We employ a keyword-element relationship summary that compactly represents relationships between keywords and the data elements mentioning them. A multilevel scoring mechanism is proposed for computing the relevance of routing plans based on scores at the level of keywords, data elements, element sets, and subgraphs that connect these elements. Experiments carried out using 150 publicly available sources on the web showed that valid plans (precision@1 of 0.92) that are highly relevant (mean reciprocal rank of 0.89) can be computed in 1 second on average on a single PC. Further, we show routing greatly helps to improve the performance of keyword search, without compromising its result quality.**

Keywords: **Keyword Query Routing, Linked Data Sources, Processing Keyword, Queries**

I. INTRODUCTION

In recent years the Web has evolved from a global information space of linked documents to one where both documents and data are linked. Underpinning this evolution is a set of best practices for publishing and connecting structured data on the Web known as Linked Data. The adoption of the Linked Data best practices has lead to the extension of the Web with a global data space connecting data from diverse domains such as people, companies, books, scientific publications, films, music, television and radio programmes, genes, proteins, drugs and clinical trials, online communities, statistical and scientific data, and reviews. This Web of Data enables new types of applications. There are generic Linked Data browsers which allow users to start browsing in one data source and then navigate along links into related data sources. There are Linked Data search engines that crawl the Web of Data by following links between data sources and provide expressive query capabilities over aggregated data, similar to how a local database is queried today. The Web of Data also opens up new possibilities for domain-specific applications. Unlike Web 2.0 mashes which work against a fixed set of data sources, Linked Data applications operate on top of an unbound, global data space. This enables them to deliver more complete answers as new data sources appear on the Web.We propose to investigate the problem of keyword query routing for keyword search over a large number of structured and Linked Data sources. Routing keywords only to relevant sources can reduce the high cost of searching for structured results that span multiple sources. To the best of our knowledge, the work presented in this paper represents the first attempt to address this problem. We use a graph-based data model to characterize individual data sources. In that model, we distinguish between an element-level data graph representing relationships between individual data elements, and a set-level data graph, which captures information about group of elements. This set-level graph essentially captures a part of the Linked Data schema on the web that is represented in RDFS, i.e., relations between classes. Often, a schema might be incomplete or simply does not exist for RDF data on the web. In such a case, a pseudo schema can be obtained by computing a structural summary such as a data guide.

1.2 Motivation

In today's world we access the Web for many needs. The Web is a collection of Linked data spread over different sources. If a user searches the Web with a simple keyword, it searches for the same across different sources and produces a large number of suggestions, of which many are not relevant to the users need. This process also implies a lot of cost in terms of time and searching. If we build a proper keyword query routing mechanism, we can improve the response time of the query and eliminate most of the suggestions that are not relevant to the keyword. shows the survey on information level required for different categories of people on a search engine . The type of information required for a graduate student on a keyword varies with information required by an under graduate student on the same keyword. There are millions of users around the world who search the Web for relevant data. They need an efficient and quickly responding search engine that can satisfy their requirements.

II. SYSTEM ANALYSIS

2.1 Existing System

Existing work can be categorized into two main categories as schema-based approaches and Schema-agnostic approaches. There are schema-based approaches implemented on top of off-the-shelf databases. A keyword query is processed by mapping keywords to elements of the database (called keyword elements). Then, using the schema, valid join sequences are derived, which are then employed to join ("connect") the computed keyword elements to form so called candidate networks representing possible results to the keyword query. Schema-agnostic approaches operate directly on the data. Structured results are computed by exploring the underlying data graph. The goal is to find structures in the data called Steiner trees (Steiner graphs in general), which connect keyword elements. Various kinds of algorithms have been proposed for the efficient exploration of keyword search results over data graphs, which might be very large. Examples are bidirectional search and dynamic programming Existing work on keyword search relies on an element-level model (i.e., data graphs) to compute keyword query results.

2.2 Disadvantages of existing system

- ❖ The number of potential results may increase exponentially with the number of sources and links between them. Yet, most of the results may be not necessary especially when they are not relevant to the user.
- ❖ The routing problem, we need to compute results capturing specific elements at the data level.
- ❖ Routing keywords return all the source which may or may not be the relevant sources

2.3 proposed system

We propose to route keywords only to relevant sources to reduce the high cost of processing keyword search queries over all sources. We propose a novel method for computing top-k routing plans based on their potentials to contain results for a given keyword query. We employ a keyword-element relationship summary that compactly represents relationships between keywords and the data elements mentioning them. A multilevel scoring mechanism is proposed for computing the relevance of routing plans based on scores at the level of keywords, data elements, element sets, and subgraphs that connect these elements. We propose to investigate the problem of keyword query routing for keyword search over a large number of structured and Linked Data sources.

2.4 Advantages of Proposed System

- ❖ Routing keywords only to relevant sources can reduce the high cost of searching for structured results that span multiple sources.
- ❖ The routing plans, produced can be used to compute results from multiple sources.

2.5 Architecture

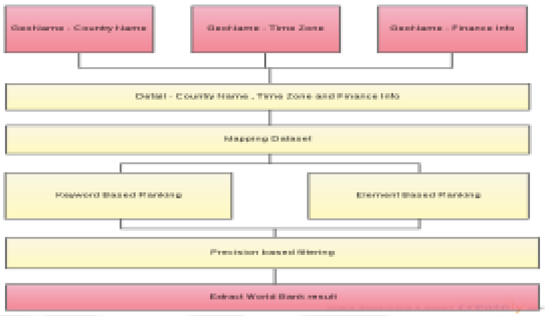


Fig 1 : Architecture

III. SYSTEM DESIGN

Linked data describes a method of publishing structured data so that it can be interlinked and become more useful. Keyword search is an intuitive paradigm for searching linked data sources on the web. We propose to route keywords only to relevant sources to reduce the high cost of processing keyword search queries over all sources. In this we have implement TOP K-Routing plan based on their potentials to contain results for a given keyword query.

Modules description:

Linked Data Generation:

The GeoNames Services makes it possible to add geospatial semantic information to the Word Wide Web. All over 6.2 million geonames toponyms now have a unique URL with a corresponding XML web service. In this we have used Country Info , Time zone and Finance Info services. This model resembles RDF data where entities stand for some RDF resources, data values stand for RDF literals, and relations and attributes correspond to RDF triples. While it is primarily used to model RDF Linked Data on the web, such a graph model is sufficiently general to capture XML and relational data.

Key level Mapping:

The set-level graph essentially captures a part of the Linked Data schema on the web that is represented in RDFS, i.e., relations between classes. Often, a schema might be incomplete or simply does not exist for RDF data on the web. In such a case, a pseudoschema can be obtained by computing a structural summary such as a data guide. A set-level data graph can be derived from a given schema or a generated pseudoschema. The web of data is modeled as a web graph where GA is the set of all data graphs, N is the set of all nodes, E is the set of all "internal" edges that connect elements within a particular source.

Multilevel Inter relationship:

The search space of keyword query routing using a multilevel inter-relationship graph. The inter-relationships between elements at different levels keyword is mentioned in some entity descriptions at the element level. Entities at the element level are associated with a set-level element via type. A set-level element is contained in a source. There is an edge between two keywords if two elements at the element level mentioning these keywords are connected via a path. We propose a ranking scheme that deals with relevance at many levels.

Routing Plan:

Given the web graph W =(G,N,E) and a keyword query K, the mapping: K-2G that associates a query with a set of data graphs is called a keyword routing plan RP. A plan RP is considered valid w.r.t. K when the union set of its data graphs contains a result for K. The problem of keyword query routing is to find the top-k keyword routing plans based on their relevance to a query. A relevant plan should correspond to the information need as intended by the user.

IV. FUNCTIONALITY

4.1 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 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 and Methods for preparing input validations and steps to follow when error occur.

4.2 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

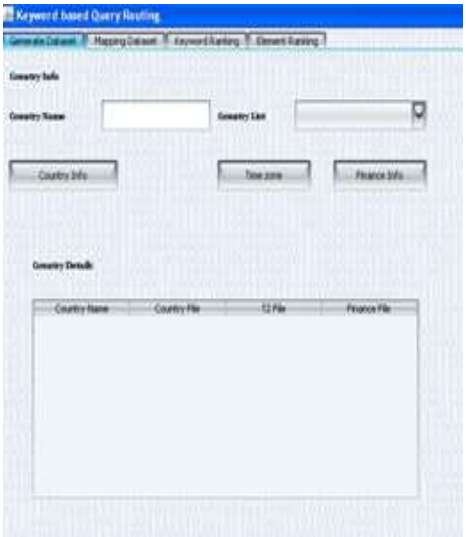
4.3 Output Design

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.

1. 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.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

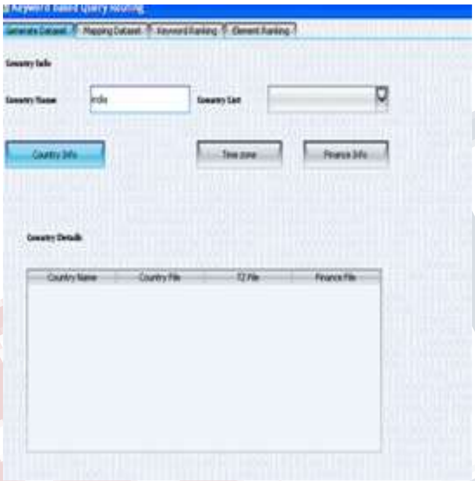
The output form of an information system should accomplish one or more of the following objectives. Convey information about past activities, current status or projections of the future, Signal important events, opportunities, problems, or warnings, Trigger an action, Confirm an action.

V. IMPLEMENTATION



Country Name	Country File	TZ File	Finance File

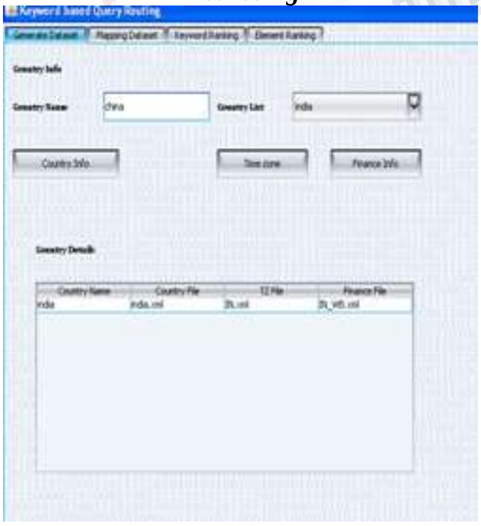
Fig 2: Login Page



Country Name: india

Country Name	Country File	TZ File	Finance File

Fig 3: The page for country information Retrieving

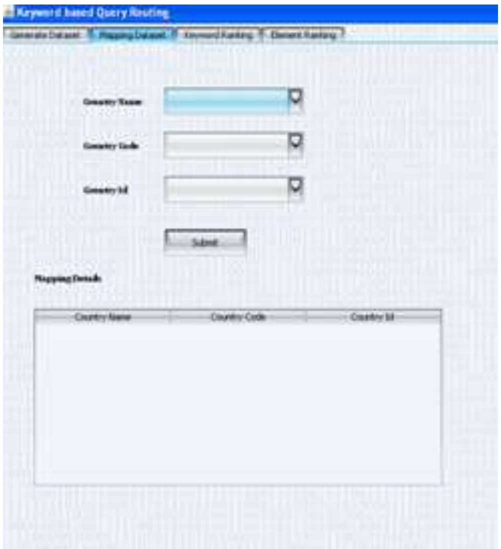


Country Name: china

Country List: india

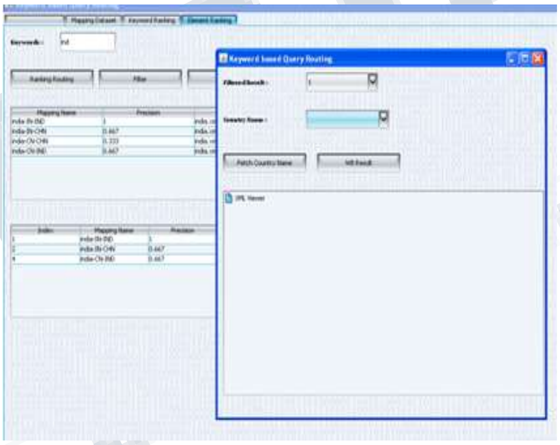
Country Name	Country File	TZ File	Finance File
india	india.xml	IN.xml	IN_WB.xml

Fig 4: The page for country Info



Country Name	Country Code	Country Id

Fig 5: The page for Mapping details



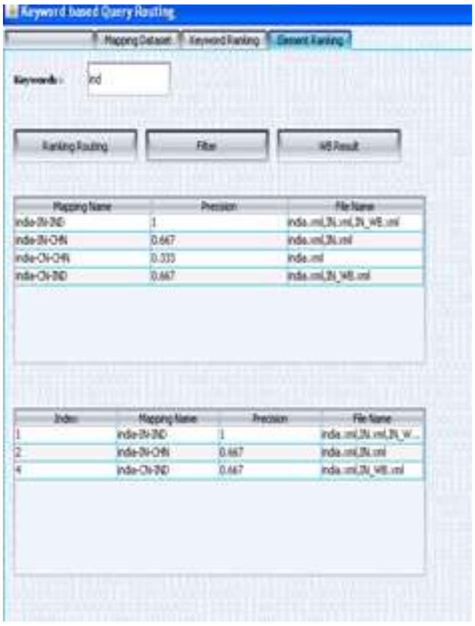
Keywords : ind

Filtered Result : 1

Mapping Name	Precision
india-IN-IND	1
india-IN-CHN	0.667
india-CN-CHN	0.333
india-CN-IND	0.667

Index	Mapping Name	Precision
1	india-IN-IND	1
2	india-IN-CHN	0.667
4	india-CN-IND	0.667

Fig 6: Window Showing the Country Information ofter Filtering



Keywords : ind

Mapping Name	Precision	File Name
india-IN-IND	1	india.xml,IN.xml,IN_WB.xml
india-IN-CHN	0.667	india.xml,IN.xml
india-CN-CHN	0.333	india.xml
india-CN-IND	0.667	india.xml,IN_WB.xml

Index	Mapping Name	Precision	File Name
1	india-IN-IND	1	india.xml,IN.xml,IN_W...
2	india-IN-CHN	0.667	india.xml,IN.xml
4	india-CN-IND	0.667	india.xml,IN_WB.xml

Fig 7: The Window Showing Mapping of Dataset

VI. CONCLUSION

We have presented a solution to the novel problem of keyword query routing. Based on modeling the search space as a multilevel inter-relationship graph, we proposed a summary model that groups keyword and element relationships at the level of sets, and developed a multilevel ranking scheme to incorporate relevance at different dimensions. The experiments showed that the summary model compactly preserves relevant information. In combination with the proposed ranking, valid plans (precision@1 ¼ 0:92) that are highly relevant (mean reciprocal rank ¼ 0:86) could be computed in 1 s on average. Further, we show that when routing is applied to an existing keyword search system to prune sources, substantial performance gain can be achieved.

REFERENCES

[1] V. Hristidis, L. Gravano, and Y. Papakonstantinou, "Efficient IR-Style Keyword Search over Relational Databases," Proc. 29th Int'l Conf. Very Large Data Bases (VLDB), pp. 850-861, 2003.

[2] F. Liu, C.T. Yu, W. Meng, and A. Chowdhury, "Effective Keyword Search in Relational Databases," Proc. ACM SIGMOD Conf., pp. 563-574, 2006

[3] Y. Luo, X. Lin, W. Wang, and X. Zhou, "Spark: Top-K Keyword Query in Relational Databases," Proc. ACM SIGMOD Conf., pp. 115-126, 2007.

[4] M. Sayyadian, H. LeKhac, A. Doan, and L. Gravano, "Efficient Keyword Search Across Heterogeneous Relational Databases," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 346-355, 2007.

[5] B. Ding, J.X. Yu, S. Wang, L. Qin, X. Zhang, and X. Lin, "Finding Top-K Min-Cost Connected Trees in Databases," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 836-845, 2007.

[6] B. Yu, G. Li, K.R. Sollins, and A.K.H. Tung, "Effective Keyword- Based Selection of Relational Databases," Proc. ACM SIGMOD Conf., pp. 139-150, 2007

[7] Q.H. Vu, B.C. Ooi, D. Papadias, and A.K.H. Tung, "A Graph Method for Keyword-Based Selection of the Top-K Databases," Proc. ACM SIGMOD Conf., pp. 915-926, 2008

[8] V. Hristidis and Y. Papakonstantinou, "Discover: Keyword Search in Relational Databases," Proc. 28th Int'l Conf. Very Large Data Bases (VLDB), pp. 670-681, 2002.

[9] L. Qin, J.X. Yu, and L. Chang, "Keyword Search in Databases: The Power of RDBMS," Proc. ACM SIGMOD Conf., pp. 681-694, 2009.

[10] G. Li, S. Ji, C. Li, and J. Feng, "Efficient Type-Ahead Search on Relational Data: A Tastier Approach," Proc. ACM SIGMOD Conf., pp. 695-706, 2009.