Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Logical and Effective Query Routing

P. Venkatesh
Department of Computer Science and Engineering
NS Raju Institute of Technology,
Visakhapatnam, Andhra Pradesh-531173, India.

K. Srikar
Department of Computer Science and Engineering
NS Raju Institute of Technology,
Visakhapatnam, Andhra Pradesh-531173, India.

P. Harika Lakshmi
Department of Computer Science and Engineering
NS Raju Institute of Technology,
Visakhapatnam, Andhra Pradesh-531173, India.

T.V.S Sriram
Department of Computer Science and Engineering
NS Raju Institute of Technology,
Visakhapatnam, Andhra Pradesh-531173, India.

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 sub graphs 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.

Index Terms: ***Keyword search, keyword query, keyword query routing, graph-structured data, RDF***

1. Introduction

THE web is no longer only a collection of textual documents but also a web of interlinked data sources (e.g., Linked Data). One prominent project that largely contributes to this development is Linking Open Data. Through this project, a large amount of legacy data have been transformed to RDF, linked with other sources, and published as Linked Data. Collectively, Linked Data comprise hundreds of sources containing billions of RDF triples, which are connected by millions of links (see LOD Cloud illustration at http://linkeddata.org/). While different kinds of links can be established, the ones frequently published are same As links, which denote that two RDF resources represent the same real-world object. A sample of Linked Data on the web is illustrated in Fig. 1. It is difficult for the typical web users to exploit this web data by means of structured queries using languages like SQL or SPARQL. To this end, keyword search has proven to be intuitive. As opposed to structured queries, no knowledge of the query language, the schema or the underlying data are needed. In database research, solutions have been proposed, which given a keyword query, retrieve the most relevant structured results [1], [2], [3], [4], [5], or simply, select the single most relevant databases [6], [7]. However, these approaches are single-source solutions. They are not directly applicable to the web of Linked Data, where results are not bounded by a single source but might encompass several Linked Data sources. As opposed to the source selection problem [6], [7], which

Cite this article as: P. Venkatesh, K. Srikar, P. Harika Lakshmi & T.V.S Sriram, "Logical and Effective Query Routing", International Journal of Research in Advanced Computer Science Engineering, Volume 4 Issue 10, 2019, Page 6-13.

Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

is focusing on computing the most relevant sources, the problem here is to compute the most relevant combinations of sources. The goal is to produce routing plans, which can be used to compute results from multiple sources.

To this end, we provide the following contributions:

- 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.
- Existing work uses keyword relationships (KR) collected individually for single databases [8-12]. We represent relationships between keywords as well as those between data elements. They are constructed for the entire collection of linked sources, and then grouped as elements of a compact summary called the set-level keyword-element relationship graph (KERG). Summarizing relationships is essential for addressing the scalability requirement of the Linked Data web scenario.
- IR-style ranking has been proposed to incorporate relevance at the level of keywords [7]. To cope with the increased keyword ambiguity in the web setting, we employ a multilevel relevance model, where elements to be considered are keywords, entities mentioning these keywords, corresponding sets of entities, relationships between elements of the same level, and inter-relationships between elements of different levels.
- We implemented the approach and evaluated it in a real-world setting using more than 150 publicly available data sets. The results show the applicability of this approach: valid plans (precision@1 ¼ 0.92) that are highly relevant to the user information need (mean reciprocal rank (RR) ¼ 0.86) can be computed in 1 second on average using a commodity PC. Further, we show that when routing is applied to an existing keyword search system to prune sources, substantial performance gain can be achieved.

2. Implementation modules

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.

2.1 Linked Data Generation

The Geo Names Services makes it possible to add geospatial semantic information to the Word Wide Web. All over 6.2 million geo names 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.

2.2 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 pseudo schema 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 pseudo schema. 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

Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

all "internal" edges that connect elements within a particular source.

2.3 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.

2.4 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.

2.5 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.

2.6 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.

3. Overview

In this section, we discuss the data, define the problem, and then briefly sketch the proposed solution.

Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

3.1 System Design

3.1.1 System Architecture:

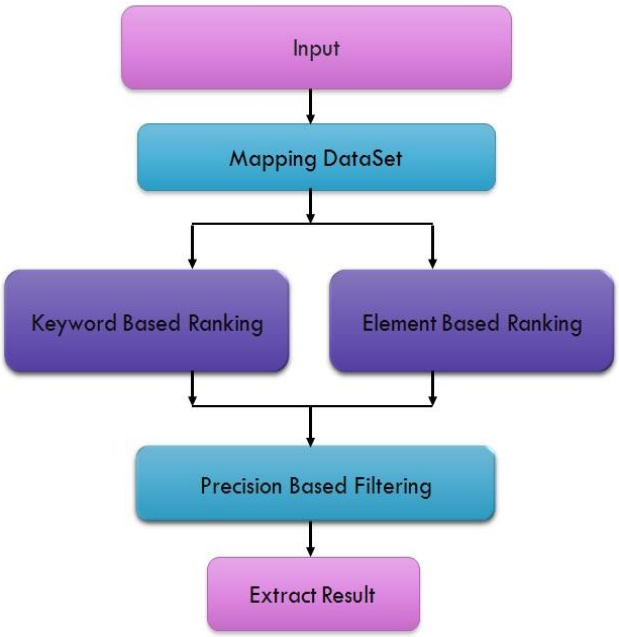


Fig. 1 System architecture

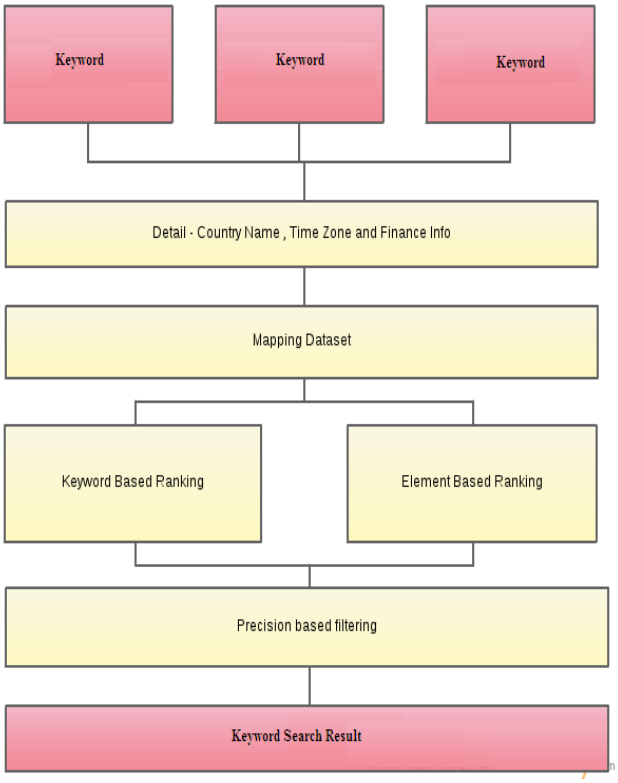
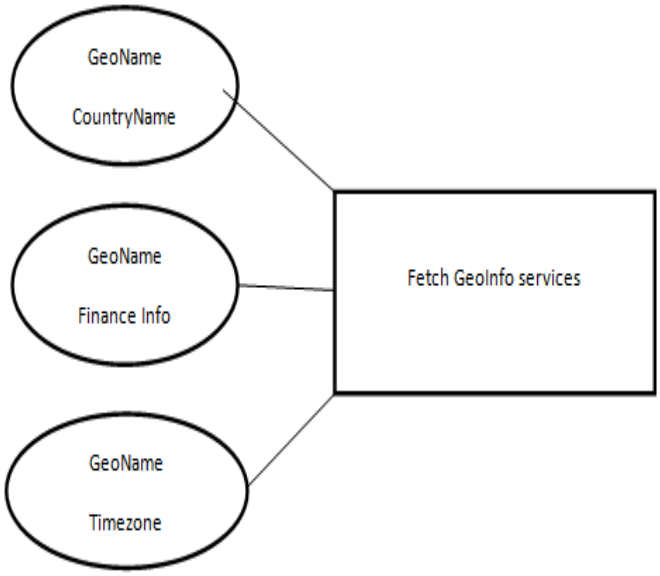


Fig. 2 System architecture example

3.2 Data Flow Diagram:

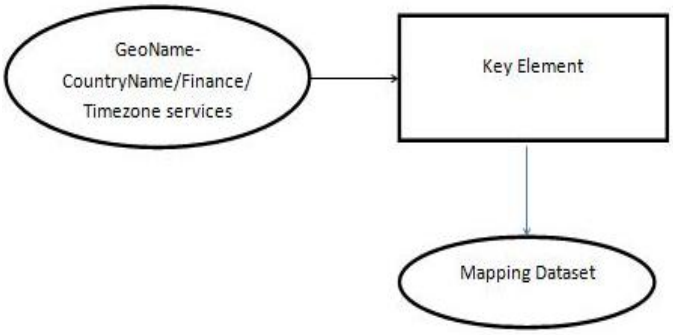
- The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
- The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
- DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
- DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

Level 0:

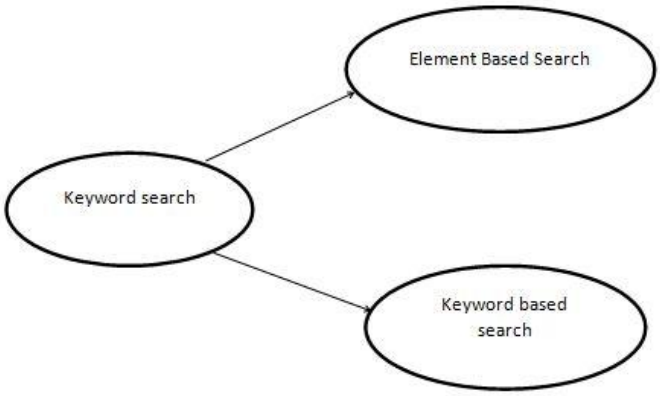


Volume No:4, Issue No:10 (March-2019)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

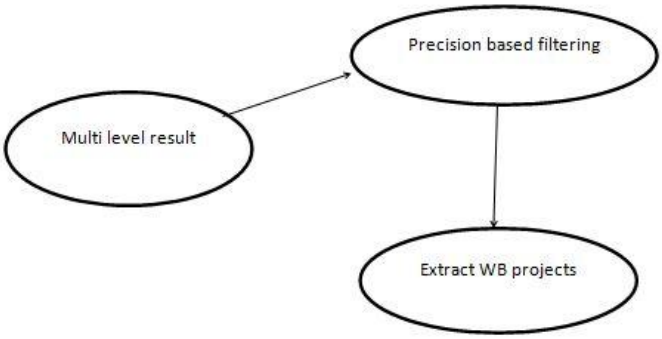
Level 1:



Level 2:



Level 3:



3.3 Activity Diagram:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

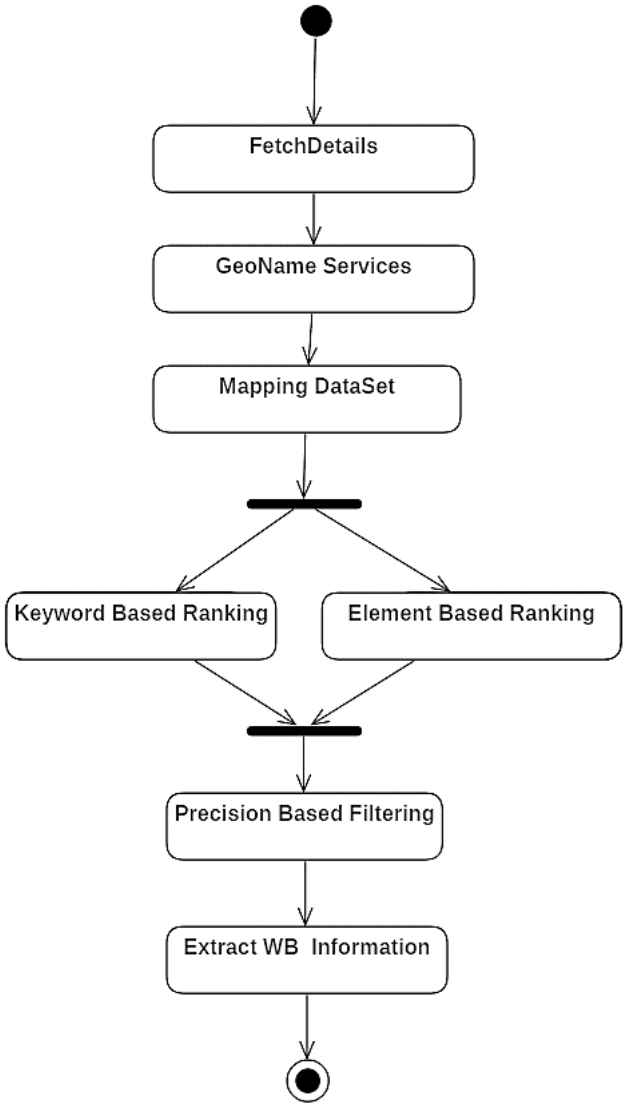


Fig. 3 Activity diagrams are graphical representations

3.4 System testing

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

Volume No:4, Issue No:10 (March-2019)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

3.4.1 Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

3.4.2 Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

4. Result and Discussions

4.1 GENERATE INFORMATION FILE RELATED TO COUNTRIES:

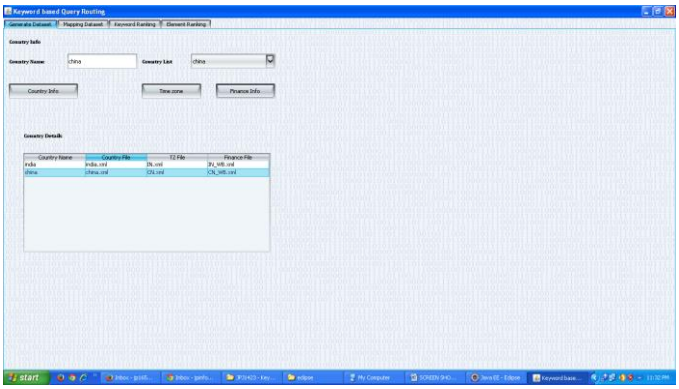


Fig 4 : Generate information file related to countries

4.2 MAPPING OF DATA SEARCH USING COUNTRY NAME, COUNTRY CODE AND COUNTRY ID:

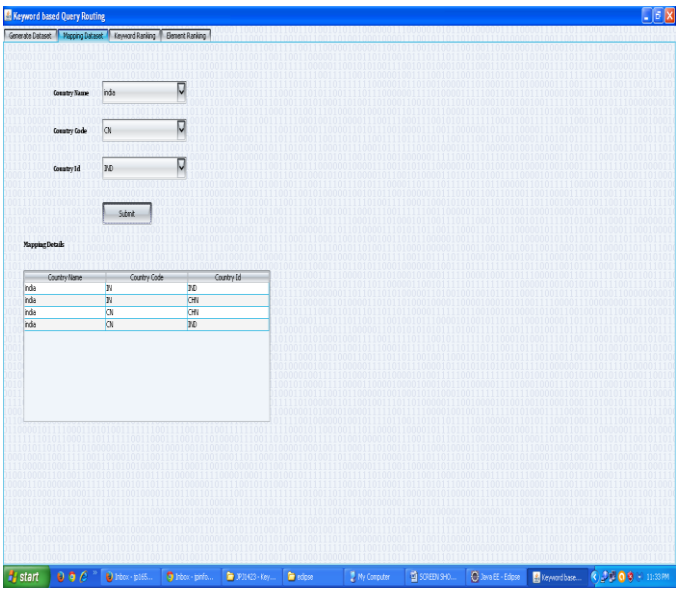


Fig 5 : Mapping of data search using country name, country code and country id

4.3 PERFORM RANKING ROUTING OPERATION RESULT WITH PRECISION VALUES USING KEYWORD RANKING:

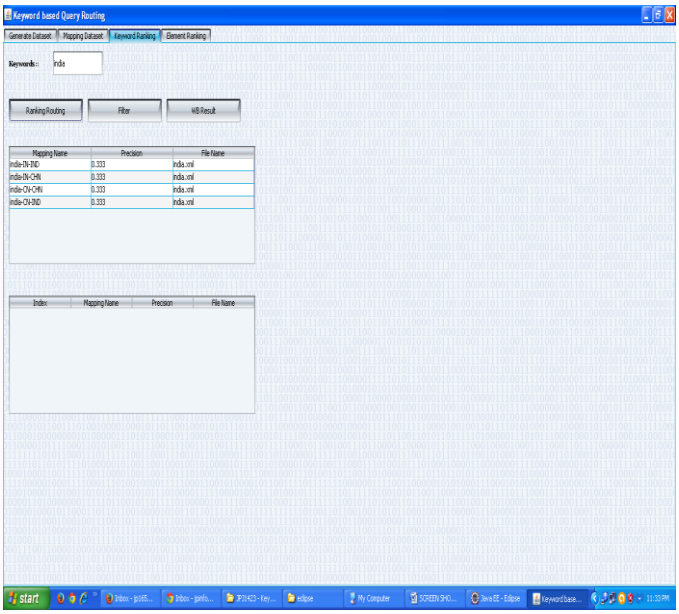


Fig. 6 Perform ranking routing operation result with precision values using keyword ranking

Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

4.4 PERFORM RANKING ROUTING OPERATION RESULT WITH PRECISION VALUES USING ELEMENT BASED RANKING:

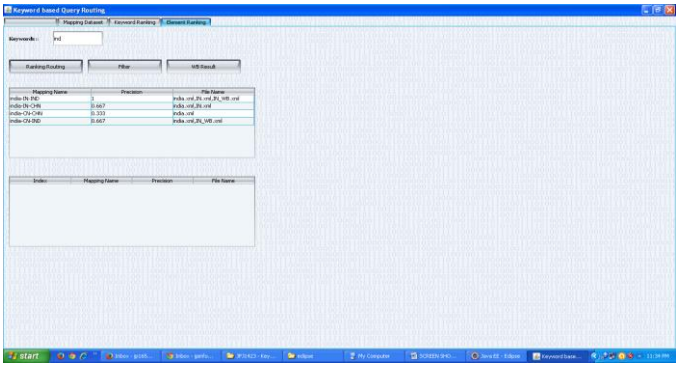


Fig. 7 Perform ranking routing operation result with precision values using element based ranking

4.5 PERFORM PRECISION BASED FILTERING OPRATION:

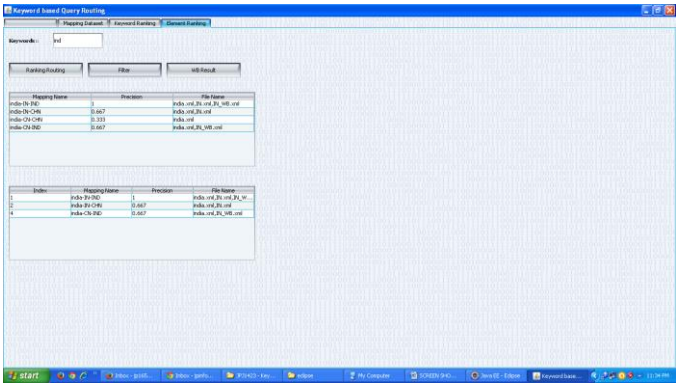


Fig. 8 Perform precision based filtering opration

4.6 ALL FINANCIAL INFORMATION FROM WB RESULTS:

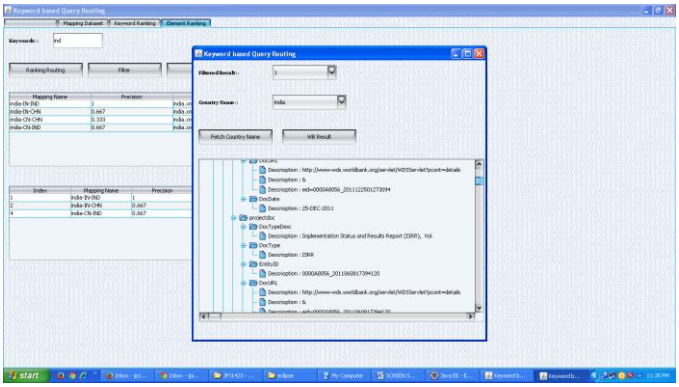


Fig. 9 All financial information from wb results

5. Conclusions

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.

6. References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

Volume No:4, Issue No:10 (March-2019)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

[6] 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.

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

[8] 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.

[9] H. He, H. Wang, J. Yang, and P.S. Yu, "Blinks: Ranked Keyword Searches on Graphs," Proc. ACM SIGMOD Conf., pp. 305-316, 2007.

[10] G. Li, B.C. Ooi, J. Feng, J. Wang, and L. Zhou, "Ease: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-Structured and Structured Data," Proc. ACM SIGMOD Conf., pp. 903-914, 2008.

[11] T. Tran, H. Wang, and P. Haase, "Hermes: Data Web Search on a Pay-as-You-Go Integration Infrastructure," J. Web Semantics, vol. 7, no. 3, pp. 189-203, 2009.

[12] G. Ladwig and T. Tran, "Index Structures and Top-K Join Algorithms for Native Keyword Search Databases," Proc. 20th ACM Int'l Conf. Information and Knowledge Management (CIKM), pp. 1505-1514, 2011.