ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

A Survey on Different approaches for Keyword Query Routing

Ashvini Gawai

M. Tech Student, Department of Computer Science and Engineering, SGGSIE&T Vishnupuri, Nanded, India

ABSTRACT: At the point when non specialized web client seeks on web it is very badly arranged for him to utilize specialized dialects like SQL to investigate the web. In request to stop this trouble keyword search has ended up being helpful. Keyword inquiry is one of the proficient approaches to recover the data of premium that might be covered up inside the different databases. To obtain the results keyword search first make a search through linked data. We work in two dimensions first one is the keyword search and the second one is database selection. In this paper we analysed keyword search approaches that is schema-based and schema-agnostic and for database selection GKS and MKS techniques are explained. One of the important terminologies for keyword search is routing plan which is the solution for keyword search**.**

KEYWORDS: Keyword query routing, MKS,GKS,RDF, Linked Data, Keyword search.

I. Introduction

Today is the era of internet. We require internet for every small and big thing. Even our smart phones will be almost useless if we don't have internet. Most of thing that is done with the help of internet is surfing which is nothing but information retrieval in technical language. Users of internet are not necessary to be a technical guy. Users of internet who don't have any knowledge of database or information retrieval also can easily surf on internet. This is possible because of number of ideas that are running in the background of this information retrieval process. One of the most popular idea behind information retrieval or surfing is Keyword Searching; means when we want some information we can simply type the word related to it and we get the desired result in no time. In Keyword Searching, result will be the databases that are relevant to the given word or query that is to be searched.

When we type any word to get information about it the process may search for this word in single database or may scan multiple databases since web consists of textual as well as non-textual documents(images, videos, different file formats etc.).The collection of this textual and non-textual document is called as Linked Data. Linked Data consists of number of data sources which may be interlinked with each other. Example of linked data is DBpedia. One of the essential terms that are included in this linked data is RDF (Resource Description Framework). RDF is nothing but the set of specifications that are given by World Wide Web Consortium (W3C).RDF is represented as directed multigraph. For specific knowledge representation entity relationship model or other ontological model are not that much useful as RDF.RDF is responsible for taking legacy data from old databases and link it with other database resources with the help of RDF triplet. RDF triplet consists of the triplet of subject, predicate and object. Linked data consists of number of RDF triplets and millions of links which are connecting these triplets. The links that are used repeatedly are called as 'sameAs' links using which we can identify real world entities. To find the answer to any query our goal is to find most relevant combinations of the databases that are used to give Routing Plans which will give answer to query by scanning multiple databases.

DOI: 10.15680/IJIRCCE.2016.0407131

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

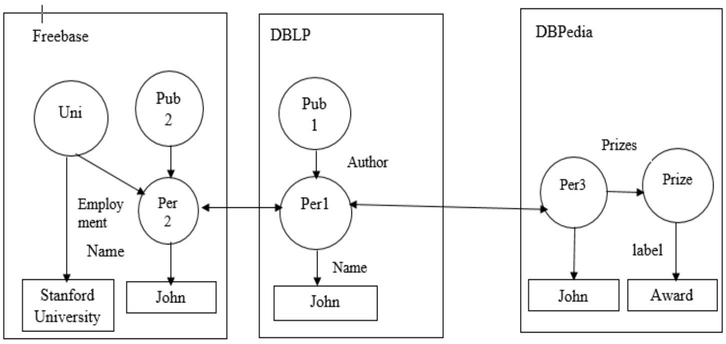

Fig: Extract of web data graph

Fig. 1.Extract of web data graph

II. Related work

Related to keyword query routing, work has been divided into two categories:
1) Keyword Search - Compute the most relevant structured results.
2) Database Selection - Compute the most relevant Sources.

A. *Keyword Search:*

Work up till now that has been done on Keyword Search can broadly classified into two classes that is

1. Schema Based approach and
2. Schema Agnostic approaches.

1. Schema Based approach

Schema based approaches takes into consideration the schemas and they operate on the top of off-the-shelf databases implies the databases which are radially accessible. A keyword query is processed by mapping the keyword in the query to the elements of the databases. Then using the schema valid join sequences are derived which are used to join computed keyword elements to form candidate networks.

Systems related to schema based approach that has been proposed before Keyword Query Routing is:
1) DBXplorer [10]
2) DISCOVER
3) BANKS
4) Hristidis

There are two stages for these systems. For the initial step, DBXplorer, DISCOVER, BANKS frameworks require an answer containing all the keywords of the query. However there is no such compulsion for Hristidis [3] system, if there are some keywords matching its alright with it. In second step, DBXplorer and DISCOVER has a very simple ranking strategy. Answers containing less number of joins are positioned before answers having more number of joins.

In case of answers having same number of joins, they are ranked arbitrarily. Due to this strategy,Answers (tuple trees) which don't contain any joins are given positioning before the answers containing joins.

Ranking strategy of BANKS is as follows:

It takes into consideration two things, first weight of tuple which is similar to web page rank and second thing is weight of edge which measures the relationship between two tuples. Ranking strategy of BANKS, DISCOVER and DBXplorer do not take into account state-of-the-art IR style ranking which is very successful. Efficiency is given in first step. Standards are planned in such an approach to keep away from superfluous tuple in answer tuple tree.

2.Schema Agnostic approach

As the name suggests schema agnostic approaches do not require any schema information for queries. Schema agnostic approaches operate directly on the top of data.Systems for schema agnostic approaches create a response as a tree containing different tuples as nodes.Result will be the structured result which is obtained by exploring data graph.The main goal for schema agnostic approach is to find structures in the data called as Steiner trees.

Systems related to schema agnostic approach that has been proposed before Keyword Query Routing are:
1) EASE
2) BLINKS

B. *Database Selection:*

The popularity of keyword based searches over internet is increasing day by day which leads to the increase in demand of keyword search over structured database.But the research work till now focuses on keyword based searching over single database.In Spite of this, as internet is growing faster it creates distributed databases and service oriented architecture over the internet.This increasing demand needs to extend the searching from single structured database to multiple structured databases.One of the important thing while extending such searching capability is to select the most relevant databases to the given keyword query.

The goal for existing work related to database selection is to find the most relevant databases.One of the most important terms related to database selection is 'keyword relationship'.Keyword relationship is nothing but the pair of keywords that can be related via a sequence of join operations.For example,as shown in figure 1,<SRTMU,Award> is a keyword relationship as there exist a path between uni1 and prize.Every database is having keyword relationship model;and according to this it is decided whether the given database is relevant or not.It is relevant if it contain all the query keywords otherwise not.

Related to keyword relationship there are two approaches, namely G-KS [12] and M-KS [12].M-KS uses the binary relationship between the keywords. Also it captures the relationship using a matrix.The drawback of M-KS is that, when we pass the query containing more than two words, then it will not give proper results and end up with errors. Due to this,if all the keyword in the query is pairwise related but it will be discarded since there is absence of a common join sequence which binds them together.

Table I: Comparison of Database Selection methods

Graph Based Keyword Search	*Matrix Based Keyword Search*
It considers more complex relationships between keywords. This is possible due to candidate graphs	It takes into account only binary relationships between pair of keywords. Due to which it gives large number of false positives.
G-KS uses graph based method to capture relationships between the keywords.	M-KS uses matrix to capture relationship between the keywords.
It requires less space.	It requires more space.
Query Processing cost is less.	Query Processing cost is high.

The solution to the above approach is G-KS system, which takes into consideration more complex relationship between keywords. G-KS uses a graph instead of matrix as in M-KS. This graph is called as 'keyword relationship graph (KRG). In the KRG, each keyword is represented by a node and the edges between the keywords <Ki, Kj>indicates that there exists at least two tuples <Ti, Tj> which are connected with each other as Ti → Tj. On these edges the distance between the two keywords is given.

KRG are used to compute the similarity between the databases and keyword query so that during query processing phase only relevant databases are to be searched.G-KS gives more relevant data sources than M-KS because of the differences listed above. GKS, MKS both assumes that, data sources are independent and answers will be derived from single database. The KRG in G-KS its keywords, relationships between the keywords are also built for a single data source. But we want the answers which will take into account the Linked Data Collection. Solution to this problem is Keyword Query Routing.

III. EXISTING SYSTEM

A. *Kite System:*

Kite System [5] is nothing but the extension of schema based approach. It is used to find the candidate networks in heterogeneous databases consisting of multiple sources. In real practice we want the answers consisting of multiple tuple tuples from different databases which may be heterogeneous and autonomous in terms of schema and data. It uses schema matching technique to find links between the sources and structure discovery techniques to search foreign key join across heterogeneous databases.

There are two phases of operation in Kite system.

1. Offline Preprocessing
2. Online Querying

In the Offline preprocessing phase, the index builder gives the inverted IR indexes on the text attributes of the database. After this, the foreign key join finder finds the foreign key joins across the databases by using data based join discovery and schema matching methods.

In Online Querying phase, we give a keyword query Q as an input, the condensed candidate network generator applies Foreign Key joins and the IR indexes to find out the space of answers to Q as early as possible. Kite uses a combination of structure discovery and schema matching methods that empirically outperforms current join discovery algorithms. As the number of databases grows the problem of keyword search in different heterogeneous databases also increases. To deal with such multi databases, the single source database solutions are ineffective since such heterogeneous databases have several different settings. To address this problem, they have introduced Kite algorithm.

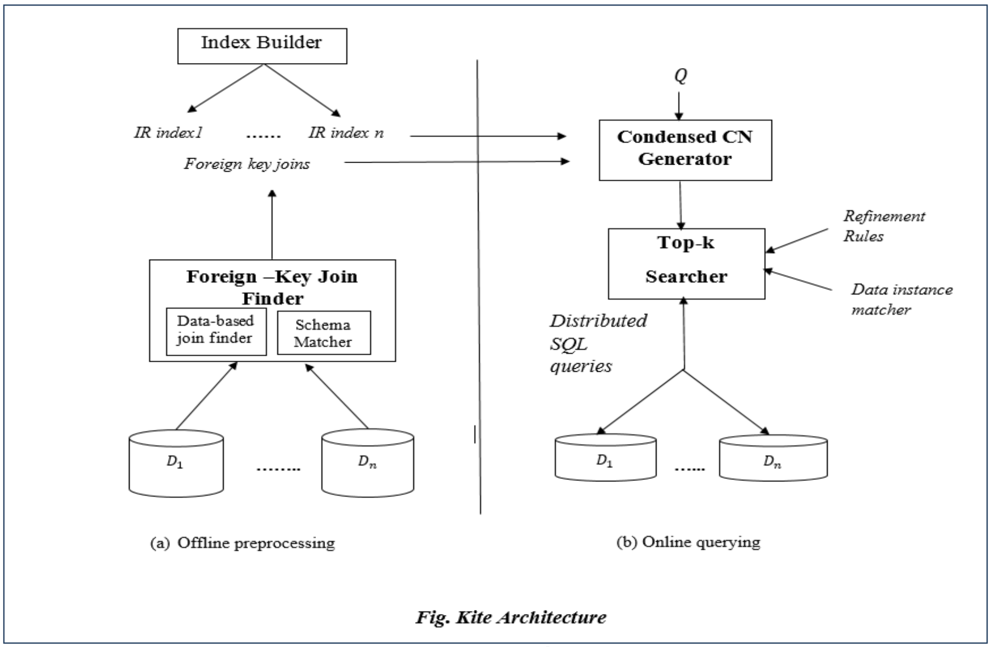

Fig. Kite Architecture

Fig.2. Kite System Architecture

Experimental studies has shown that Kite scales well with multiple databases and gives best results without any need to hide those different databases.

There are two phases of operation in Kite system.

1. Offline Preprocessing
2. Online Querying

In the Offline preprocessing phase, the index builder gives the inverted IR indexes on the text attributes of the database. After this, the foreign key join finder finds the foreign key joins across the databases by using data based join discovery and schema matching methods.

In Online Querying phase, we give a keyword query Q as an input, the condensed candidate network generator applies Foreign Key joins and the IR indexes to find out the space of answers to Q as early as possible. Kite uses a combination of structure discovery and schema matching methods that empirically outperforms current join discovery algorithms.

As the number of databases grows the problem of keyword search in different heterogeneous databases also increases. To deal with such multi databases, the single source database solutions are ineffective since such heterogeneous databases have several different settings. To address this problem, they have introduced Kite algorithm.

Experimental studies has shown that Kite scales well with multiple databases and gives best results without any need to hide those different databases.

B. *EASE(An Effective 3-In-1 Keyword Search Method for unstructured semi structured and Structured Data):*

If we have a large collection of heterogeneous data then it is very difficult to query and index such data. This paper proposes a very efficient adaptive method for keyword search over large heterogeneous database called ease. For keyword search they proposed an extended inverted index, and novel ranking mechanism for enhancing search effectiveness. After making particle experiments they showed that ease achieves both search efficiency and accuracy which outperforms existing approaches.

C. *BLINKS:*

When we pass a top-k keyword search query on graph database then it will retrieve the answers containing some of the portions of the graph means some tuples of the graph which contain all the keywords of the given query.

BLINKS [2] is a bi level indexing and query processing scheme for top-k keyword searching on graphs. BLINKS have a search strategy which has performance bounds and it exploits bi-level index for pruning and speed up the search. Since it is bi-level index, and to speed up the search, it divides the data graph into blocks. For these blocks, the bi-level index stores the summary information which initiate guide search among blocks and gives more detail information about blocks to speed up search among blocks.

The main contribution of BLINKS system is:

1. Better search strategy.
2. Combining indexing with search.
3. Partitioning based indexing.

IV. COMPARISON BETWEEN DIFFERENT KEYWORD SEARCH APPROACHES

There are two stages for these systems. For the initial step, DBXplorer, DISCOVER, BANKS frameworks require an answer containing all the keywords of the query. However there is no such impulse for Hristidis system, if there are some keywords matching its alright with it. In second step, DBXplorer and DISCOVER has a very simple ranking strategy. Answers containing less number of joins are positioned before answers having more number of joins.In case of answers having same number of joins, they are ranked arbitrarily. Due to this strategy,Answers (tuple trees) which don't contain any joins are given positioning before the answers containing joins.

Ranking strategy of BANKS is as follows:

It takes into consideration two things, first weight of tuple which is similar to web page rank and second thing is weight of edge which measures the relationship between two tuples. Ranking strategy of BANKS, DISCOVER and DBXplorer do not take into account state-of-the-art IR style ranking which is very successful. Efficiency is given in first step. Standards are planned in such an approach to keep away from superfluous tuple in answer tuple tree.

DOI: 10.15680/IJIRCCE.2016.0407131

Table II: Comparison of Database Selection methods

Working Strategy	**Existing Systems**			
	DBXplorer	***DISCOVER***	***BANKS***	***Hristidis***
Answer contain all keywords of query in 1 st stage	Yes	Yes	Yes	No
Ranking Strategy	Simple	Simple	Complex	Complex
Support State-of-art IR ranking strategy	No	No	No	Yes

V. Conclusion and Future Work

In this survey paper, I have analysed different approaches for keyword query routing. This analysis includes the detail study of keyword search techniques and database selection techniques. For the keyword search there are two approaches schema based approach and schema agnostic approach. Also for database selection there are two approaches GKS (Graph Based Keyword Search) and MKS (Matrix Based Keyword Search).

The keyword search approaches gives the most relevant structured results while database selection techniques gives the most relevant data source which solves the problem of source selection. The proposed system gives a new system for keyword search which makes use of routing plans. This system route keywords to only relevant sources so as to reduce the high cost of processing the given keyword query over all the sources which further reduces the processing time when compared to previous approaches.

In future work, an approach can be developed where there will be a ranking technique for linked databases which can reduce the query execution time. An approach can be developed for an efficient algorithm for fuzzy keyword search. Moreover, one can design an algorithm for such datasets where there is no common attributes.

References

1. Thanh Tran and Lei Zhang, "Keyword Query Routing", IEEE Transactions On Knowledge And Data Engineering, VOL. 26, NO. 2, FEBRUARY 2014
2. 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.
3. 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.
4. F. Liu, C.T. Yu, W. Meng, and A. Chowdhury, "Effective Keyword Search in Relational Databases", Proc. ACM SIGMOD Conf., pp. 563-574, 2006.
5. 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.
6. V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, and H. Karambelkar, "Bidirectional Expansion for Keyword Search on Graph Databases", Proc. 31st Int'l Conf. Very Large Data Bases (VLDB), pp. 505-516, 2005.
7. 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.
8. H. He, H. Wang, J. Yang, and P.S. Yu, "Blinks: Ranked Keyword Searches on Graphs", Proc. ACM SIGMOD Conf., pp. 305-316, 2007
9. 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.
10. S. Agrawal, S. Chaudhuri, and G. Das, "DBXplorer: A system for keyword-based search over relational databases",In ICDE, 2002
11. 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.
12. 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.

BIOGRAPHY

Ashvini B. Gawai is student of Computer Network and Information Security (CNIS) Department of Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded. She received Master of Technology (M.Tech) degree in 2016 from SRTMU, Nanded, MS, India. Her research interests are Big Data and Data Mining.

DOI: 10.15680/IJIRCCE.2016.0407131