ISSN 2395-1621

A Survey Paper on Keyword Search Mechanism for RDF Graph Model

[#1]Manisha Bhaik,[#2]Shyam Gadekar,[#3]Nikhil Gumaste,[#4]Laxmikant Suryawanshi

[1]manishabhaik1@gmail.com,
[2]shyam.gadekar123@gmail.com,
[3]nsgumaste@gmail.com,
[4]laxmikants184@gmail.com

[#1234]Department of Computer Engineering
JSPM's
Imperial College of Engineering & Research
Wagholi, Pune-412207

ABSTRACT

We present the implementation of a keyword based querying system operating on RDF databases. As in various search technique keyword is used which provides a simple but user friendly interface to retrieve information from complicated data structure. Most of these knowledge bases adopt the Semantic-Web data model RDF as a representation model. Querying these information bases is classically done using structured queries utilizing graph-pattern languages like as SPARQL. However, queries require some expertise from users which limits the accessibility to such data sources. To overcome this drawback, keyword search will be supported. This paper used indexing, pruning and refinement phases. This method provides efficient result for searching keyword on graph. Approximate mining algorithms can be used to form sub graph from RDF graph data based on scores at the level of keywords, data elements, element sets, and sub graphs that join these elements. To retrieve the well-organized keyword from sub graph keyword matching algorithm can be used for graph data. The purpose of this technique is to reduce the high cost of processing keyword search queries on graph information and get better performance of keyword search, without compromising its result quality. Also, it reduces processing time for keyword search in RDF graph data.

ARTICLE INFO

Article History

Received:28th September 2015

Received in revised form :

1st October 2015

Accepted:5th October, 2015

Published online :

6th October 2015

***Keywords*:-Data Mining, RDF Graph, Semantic Web, SPARQL, Keyword Search.**

I. INTRODUCTION

In various real world applications, RDF (Resource Description Framework) has been widely used as a W3C standard to describe data in the Semantic Web. RDF data may often suffer from the unreliability of their data sources, and exhibit irregularity or errors. In this paper, we model such unreliable RDF data by probabilistic RDF graphs [1] and learn a vital problem; keyword search query over probabilistic RDF graphs (i.e. the pg-KWS query). To retrieve meaningful keyword search results, we implement the score rankings for sub-graph answers specific for RDF data.

Now a day, keyword search is the leading technique of searching a data source such as the Web. Using solely keywords, i.e., a small number of highly perceptive terms the consumer anticipates that she will identify the web pages most relevant to her information needs. Keyword search offers a straightforward, intuitive, and yet flexible method of retrieving information. The success of keyword search in the field of Information Retrieval (IR) and the World Wide Web (WWW or just Web) has generated awareness in keyword search interacts to relational databases and similar structured and semi structured data sources. This is the way how keyword search has evolved over time and it has been adopted by different fields in computer science, such as IR, databases, and semantic web, and surveys the state of the art in keyword search for fields managing structured and semi structured data. Beyond that, it presents and extensively evaluates the design and implementation of a system working on RDF data, accommodating keyword queries with temporal constraints. At last, it presents the state of the

art in assessment of keyword search system working on structured and semi structured data.

Query processing over graph data has attracted considerable attention recently as an increasing amount of data which is available on the web, XML data sources and relational sources can be modelled in the form of graphs. RDF as a framework for web resource description appears to have gained a larger impetus on the web and an increasing collection of repositories of data are modelled using RDF framework.

Notable examples are Biological Databases, Personal Information Systems where e-mails, papers and images are merged into a graph and Enterprise Information Management (EIM) systems like launch vehicle blueprint information where details of vehicle parameters and stage sequence events is modelled as graph data. The large size and complication of data sets in these domains makes their querying a difficult job.

The keyword search over RDF graph is useful in applications in Semantic Web. RDF graph consist of RDF resources, RDF schema, and their vertices related to information of keyword. SPARQL is standard language of RDF graph. In semantic web, during data extraction/integration, data contains errors or a problem of data inconsistency because of data contains irregular format or unstructured texts. Also, there are various types of information extraction methods.

Because of unreliability of data, we integrate with RDF graphs and keyword search becomes efficient. Therefore we call the RDF graph as also probabilistic graph.

Example: YAGO data set [20] which contains probabilistic RDF data integrated from WordNet and Wikipedia. RDF triples which are (*subject, predict, object*) or (S.P.O.).

II. RELATED WORK

RDF has been used in the data mining for accurate keyword search with the help of the different data schema of RDF data, and improves the performance throughout the process in the lifecycle.

SPARQL query is a standard SQL-like query language for RDF data. For knowing the SPARQL query, have to know schema of RDF data, including the subject, predicate, or object. It has used several data models, like as triple store [4], [17], [23], Column-store [2], [18], [19], property tables, [24], [25], and graphs [3], [21].

Works held previously have the problems of efficiency and processing on the data. Already, there existing works on searching methods like as r-Radius Graph from EASE [12] keyword Search Method for all type of data. From the works IR-based ranking score functions [21], [1] like as matching and popularity score.

Probabilistic RDF databases works done by graph representation relation by Fukushige[7] in Bayesian network. The work done in the SPARQL [9] queries as an alternative of keyword search which will provide high flexibility and performance in Lian and Chen [15] has the efficient query answering with RDF Schema.

Several probabilistic queries for unstructured data have projected, as a probabilistic range query (PRQ) [6], nearest neighbour (PNN) [5], and reverse nearest neighbour (PRNN) [14].

Top-k queries [13] gives retrieve tuples with variety of probabilities and score. Keyword Search with Graph gives more ease with searching throughout the databases. Existing works provides us significant query keyword, and different levels of abstraction from graphs. Tree with root r [8], [11] and other leaf node contains the query keywords. Ranking Score has been calculated with path length. BANKS [10] has backward search method, which will traverse thoroughly with link with predecessor.

Above, all works states that there have keyword search over graphs.

III. EXISTING SYSTEM

A Resource Description Framework (RDF) is a W3C standard to represent resource information in the semantic web. The keyword search over graphs has drawn much awareness from the database community due to many applications. In this function, the original data are often represented by graph structures, in which vertices/edges are associated with keywords. In the RDF graphs is also useful in real application such as searching the Semantic Web with query keywords.

While searching queries on the graphs, RDF graph database has graph G which modelled by Bayesian networks [22].

A probabilistic RDF graph G [1], [22] is a triple {V(G), E(G), S(G)}:

V(G) is a set of vertices v_i

E(G) is a set of directed edges e_i

S(G) is a set of conditional probability table (CPT).

From the base paper [1] we get that uncertain keywords in vertices, and certain labels in edges. Edge labels are uncertain we get extended solution for the RDF graph.

It has also we add virtual vertex to each edge and make probabilistic RDF graph and assign uncertain edge keyword to RDF graph.

There have RDF Schema which will store real RDF data like as elements and knowledge base structure.

The Keyword Search Query over Probabilistic RDF Graph, (pg-KWS).

There are two keyword searching methods [1] were used. They are as follows:

a. The *r*-Radius Graph

b. Ranking Scores for RDF Keyword Search

r-Radius graph g gives smallest unit of data which uses Dijkstra Algorithm. The ranking score for keyword search uses IR-based score [16], structural score [8], [11], and combination of both score types [12].

From the base paper [1], we get Equation as follows.

$$P(g) = P_r \{g \text{ is not dominated by } g' | \forall g'\}$$

$$\sum_{\forall pw(g) \in PW(g)} \Pr\{pw(g)\} \cdot Pr\left\{ \bigwedge_{\forall g'} g' \prec pw(g) \right\}$$

Eq. (1)

IV. PROPOSED SYSTEM

To resolution confliction in rdf data, we make rdf graph as a probabilistic graph. in particular, keywords of vertices in probabilistic rdf graphs are allocated with probabilities, indicating the confidences that keywords are (conditionally) true in reality, which can be inferred from the reliability of different data sources. we propose the pg-kws problem in probabilistic rdf graphs with suitable ranks. we design effective pruning strategies to reduce the pg-kws search space. we propose an efficient approach to process pg-kws queries via pre-computed. we conduct extensive experiments to show the pg-kws query performance. this paper will undertake the keyword search problem on probabilistic rdf graphs (namely, pg-kws).

Our proposed system will reduce high cost processing keyword search queries on graph data, improve performance of keyword search and it will reduce processing time for search keyword.

- **Implemented framework:**

In particular, the framework consists of three phases, indexing, pruning, and refinement phases. In the indexing phase,

1. Indexing Phase: We will offline extract probabilistic r-radius graphs from the probabilistic RDF graph, and precompute data for each graph. Then, we construct an index over these pre-computed data for probabilistic r-radius graphs, which will be used later for online pruning and pg-KWS query answering.

2. Pruning Phase: Given any pg-KWS query, the second pruning phase traverses the index, and meanwhile applies pruning methods to quickly rule out false errors (i.e. those sub graphs that cannot be pg-KWS query answers). In particular, we will intend two pruning strategies, Score bound pruning and Probabilistic pruning, which utilize score bounds or probabilistic threshold, respectively, to enable the pruning. After the index traversal, we can obtain a candidate set Scand.

3. Refinement Phase: Finally, in the refinement phase, we refine candidates in Scand by checking the condition in Eq. (1), and return the actual pg-KWS answers.
    a. We will use following methods for pg-KWS Query Processing,
    b. Index Construction
    c. Pruning with Index Nodes
    d. Query Procedure

Above works states till there have some enhancement required in it.

- **SYSTEM ARCHITECTURE**

In the system architecture, there has two types of server maintained which are Normal Database Server and RDF Graph Database Server. Normal Data Server which is

Traditional they are used in the daily information storage and operations. RDF Graph Server stores the processed information with the help of RDF Framework which is shown in Fig.1.

The unstructured or structured database contains raw data or knowledge base information which helps to make it in RDF Triple format. It has mainly three phases which are as follows:

i. Indexing Phase
ii. Pruning Phase
iii. Refinement Phase

Above all three phases has been explained in Section IV. The indexing extract probabilistic r-radius graphs from the probabilistic RDF graph and send it to the pruning phase.

Then, the pruning phase traverses the index and applies the pruning methods for rule out the false errors or notifications. After all the data has sent to the refinement phase for purifying by checking condition in Eq.1 [1].

All processed data has stored in RDF Graph Database. When the query fired by user on search engine it will on the RDF Graph Server and it will display results to the user.

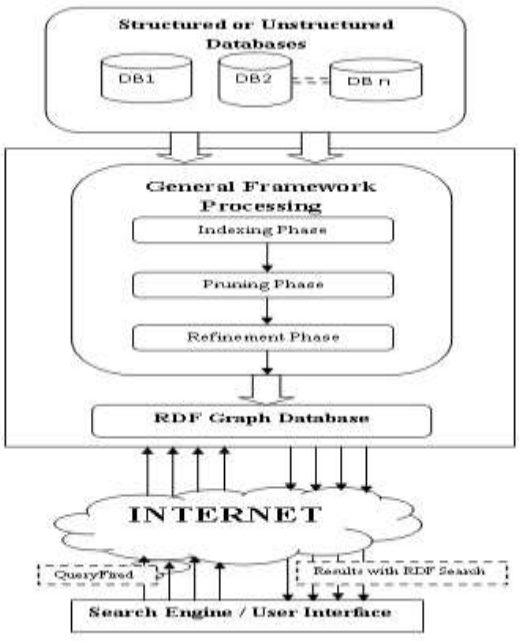

Fig.1 System Architecture

In the RDF Graph Database Server, data often represented by graph structure, in which vertices/edges are associated

with keywords. Each vertex is often associated with one or multiple keywords. Structure of the RDF Triples is as follows,

(*Subject, Predicate, Object*) or (*S. P. O.*)

Example:

We use an RDF triples (<max>, <read>, <books>) to describe a case that . We modify the triple to a directed edge as "reads", as shown below,

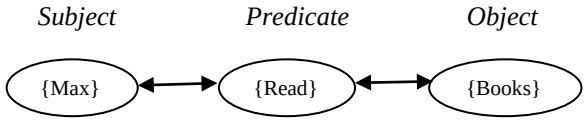

Fig.2 Graph Interconnection Structure

This will be efficient and effective mechanism for keyword searching in RDF Graph. It will help to maintain friendly relationship between edges and vertices.

V. CONCLUSION

We invent and undertake an important problem of keyword search over probabilistic RDF graphs, called pg-KWS queries. We will propose efficient pruning methods via offline pre-computed score bounds and probabilistic threshold to quickly filter out false errors. Furthermore, we construct a directory for the pre-computed data for RDF and present an efficient query resulting approach. General research has been conducted to verify the effectiveness and efficiency of our proposed approaches.

VI. ACKNOWLEDGEMENT

The authors wishes to thank Prof. Rashmi Sonawane (Guide), Prof.Darshika Lothe(PG–Coordinator), Prof. S.R.Todmal (HOD) and Dr. Sachin Admane (Principal) for valuable guidance and encouragement.

REFRENCES

[1] aXiang Lian, Lei Chen, Member, IEEE, and Zi Huang, Member, IEEE "Keyword Search Over Probabilistic RDF Graphs" in Proc.IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 5, May 2015, pp.1246-1260.

[2] D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach, "Scalable semantic web data management using vertical partitioning," in Proc. 33rd Int. Conf. Very Large Data Bases, 2007, pp. 411–422.

[3] R. Angles and C. Gutierrez, "Querying RDF data from a graph database perspective," in Proc. 2nd Eur. Conf. Semantic Web: Res. Appl., 2005, pp. 346–360.

[4] M. Atre, V. Chaoji, M. J. Zaki, and J. A. Hendler, "Matrix "bit" loaded: A scalable lightweight join query processor for RDF data," in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 41–50.

[5] G. Beskales, M. Soliman, and I. F. Ilyas, "Efficient search for the top-k probable nearest neighbors in uncertain databases," in Proc. 34th Int. Conf. Very Large Data Bases, 2008, pp. 326–339.

[6] R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Evaluating probabilistic queries over imprecise data," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2003, pp. 551–562.

[7] Y. Fukushige, "Representing probabilistic relations in RDF," ISWC-URSW, pp. 106–107, 2005.

[8] H. He, H. Wang, J. Yang, and P. S. Yu, "BLINKS: Ranked keyword searches on graphs," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2007, pp. 305–316.

[9] H. Huang and C. Liu, "Query evaluation on probabilistic RDF databases," in Proc. 10th Int. Conf. Web Inform. Syst. Eng., 2009, pp. 307–320.

[10] A. Hulgeri and C. Nakhe, "Keyword searching and browsing in databases using BANKS," in Proc. 18th Int. Conf. Data Eng., 2002, pp. 431–440.

[11] V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, and H. Karambelkar, "Bidirectional expansion for keyword search on graph databases," in Proc. 31st Int. Conf. Very Large Data Bases, 2005, pp. 505–516.

[12] 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," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2008, pp. 903–914.

[13] J. Li, B. Saha, and A. Deshpande, "A unified approach to ranking in probabilistic databases," Proc. VLDB Endowment, vol. 2, no. 1, pp. 502–513, 2009.

[14] X. Lian and L. Chen, "Efficient processing of probabilistic reverse nearest neighbor queries over uncertain data," The VLDB J., vol. 18, pp. 787–808, 2009.

[15] X. Lian and L. Chen, "Efficient query answering in probabilistic RDF graphs," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2011, pp. 157–168.

[16] Y. Luo, X. Lin, W. Wang, and X. Zhou, "Spark: Top-k keyword query in relational databases," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2007, pp. 115–126.

[17] T. Neumann and G. Weikum, "RDF-3X: A risc-style engine for RDF," Proc. VLDB Endowment, vol. 1, no. 1, pp. 647–659, 2008.

[18] L. Sidirourgos, R. Goncalves, M. Kersten, N. Nes, and S. Manegold, "Column-store support for RDF data management: Not all swans are white," Proc. VLDB Endowment, vol. 1, no. 2, 2008, pp. 1553–1563.

[19] M. Stonebraker, D. J. Abadi, A. Batkin, X.Chen, M. Cherniack, M. Ferreira, E. Lau, A. Lin, S. Madden, E. O'Neil, P. O'Neil, A. Rasin, N. Tran, and S. Zdonik, "C-store: A column-oriented dbms," in Proc. 31st Int. Conf. Very Large Data Bases, 2005, pp. 553–564.

[20] F. M. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A large ontology from wikipedia and wordnet," Web Semantic, vol. 6, no. 3, pp. 203–217, 2008.

[21] T. Tran, H. Wang, S. Rudolph, and P. Cimiano, "Top-k exploration of query candidates for efficient keyword search on graph-shaped (RDF) data," in Proc. IEEE Int. Conf. Data Eng., 2009, pp. 405–416.

[22] D. Z. Wang, E. Michelakis, M. Garofalakis, and J. Hellerstein, "Bayestore: Managing large, uncertain data repositories with probabilistic graphical models," in Proc. Int. Conf. Very Large Data Bases, 2008, pp. 340–351.

[23] C. Weiss, P. Karras, and A. Bernstein, "Hexastore: Sextuple indexing for semantic web data management," Proc. VLDB Endowment, vol. 1, no. 1, pp. 1008–1019, 2008.

[24] K. Wilkinson, "Jena property table implementation," in Proc. Scalable Semantic Web Knowl. Base Syst., 2006, pp. 35–46.

[25] K. Wilkinson, C. Sayers, H. Kuno, D. Reynolds, and J. Database, "Efficient RDF storage and retrieval in Jena2," in Proc. Eur. Semantic Web Databases workshop, 2003, pp. 131–150.

