

KEYWORD SEARCH QUERY OVER PROBABILISTIC RDF GRAPHS USING EFFECTIVE PRUNING METHODS

T.T.Mathangi,

Assistant Professor,

Department of Computer Science & Engineering,
K.L.N. College of Information Technology,
Pottapalayam, Sivagangai, Tamilnadu.

S.Dhivya,

Assistant Professor,

Department of Computer Science & Engineering,
K.L.N. College of Information Technology,
Pottapalayam, Sivagangai, Tamilnadu.

S.Jeniba,

Assistant Professor,

Department of Computer Science & Engineering,
K.L.N. College of Information Technology,
Pottapalayam, Sivagangai, Tamilnadu.

Abstract: Keyword searches are a good substitute for a subject search. It is useful when you do not know the standard subject heading. Many large-scale RDF knowledge bases with billions of facts are generating due to the growth of semantic data on the Web. This poses significant challenges for the storage and query of big RDF graphs. It is used for exploring large RDF (Resource Description Framework) Graph which is used as a W3C standard to describe data in the Semantic Web for many real time applications. In Existing techniques, RDF Data may often suffer from the unreliability of their data sources, and exhibit errors or inconsistencies. In this paper, we overcome these kinds of inconveniences with the help of an important problem, keyword search query over probabilistic RDF graphs (namely, the pg-KWS query) by probabilistic RDF graphs. To retrieve meaningful keyword search answers, we design the score rankings for sub graph answers specific for RDF data. Furthermore, we propose effective pruning methods to quickly filter out false alarms. We construct an index over the pre-computed data for RDF. We also present an efficient query answering approach through the index. Experiments have been conducted to verify the proposed approaches.

Keywords: Probabilistic RDF graph, keyword search, pg-KWS

I. INTRODUCTION

Resource Description Framework (RDF) is a W3C standard to represent resource information in the Semantic Web[1]. Nowadays, in many real applications, the RDF format has been widely used to model semantic data, in the form of subject, predicate, object. Equivalently, a collection of RDF triples can be represented by a directed graph [13] in which vertices are indicated by subjects (S) or objects (O), and edges are indicated by predicates (P). For example, we use an RDF triple (hauthori, hwrites i, hnovel_seriesi) to describe a case that an author writes a novel series. Equivalently, we can transform this triple to a directed edge (labeled by “writes”) that connects from vertex “author” to vertex “novel_series”, as shown in RDF Graph g_1 of Fig. 1a. In the literature [25], the keyword search over graphs has drawn much attention from the database community due to many applications (e.g., chemical data analysis and social network mining). In these applications, the underlying data are often represented by graph structures, in which vertices/edges are associated with keywords. In the RDF graph of Fig. 1a, each vertex is often linked with one or multiple keywords (extracted from the label or URI of the vertex), for example, the vertex that represents the character of the novel has two keywords {“John”, “Smith”}.

The keyword search is quite useful, due to its ease of use. That is, the keyword search techniques do not require users to

master the domain knowledge of the data schema (e.g., graph structures and keywords), and also it return satisfactory query results by only specifying a few enquiry keywords (rather than a detailed query graph). The keyword search is useful in real applications such as searching the Semantic Web with query keywords. Different from general graphs, RDF graphs contain RDF resources following the RDF schema, and their vertices are often associated with (imprecise) keyword information. Further, queries on general graphs include the subgraph matching and keyword search queries, which do not have to consider RDF properties (e.g., RDF resources or schema). In contrast, RDF has its own standard language, SPARQL, which requires the knowledge of RDF graph structure and labels. Moreover, keyword search queries over RDF graphs [25] need to consider properties of RDF graphs (e.g., RDF resources), which is the focus of our problem.

During the data extraction in the Semantic Web, RDF data often contain errors [15], [19]. For example, in the data extraction application, we apply information extraction (IE) methods, and extract RDF data from unstructured texts on the web. The resulting RDF data may acquire incorrectness, due to the inaccuracy of IE techniques. Moreover, with different IE methods, extraction results may exhibit inconsistencies (e.g., with conflicting keywords for the same RDF resources).

Due to the unreliability of data sources (e.g., data expirations), when we integrate RDF graphs from different sources [11], keyword information of the same RDF entity may conflict with one another. For example, Figs. 1a and 1b show two RDF graphs g_1 and g_2 , from data sources A and B, respectively. Although most portions of the two graphs are the same, there exist some different vertex keywords.

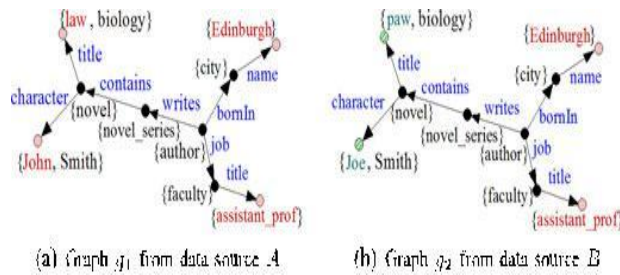


Fig 1: RDF graphs from different data sources (keywords: “John”, “law”, “Edinburgh”, and “assistant_prof”).

Thus, this paper will deal with the keyword search problem on probabilistic RDF graphs (namely, pg-KWS). That is, we want to retrieve probabilistic RDF subgraphs, which contain query keywords and have both high existence probabilities and ranks. Specifically, there are many real applications of the pg-KWS query. As an example, in the application of the job recruitment, a project manager may want to identify a team of engineers, who have close relationships and qualified skills (such as C++, UML, SQL, etc.) to assign tasks, from social networks. In this case, we can treat the integrated networks as probabilistic graphs, and issue a pg-KWS query over probabilistic graphs. That is, the project manager (user) can specify a set of query keywords, including the required skills, and retrieve the pg-KWS query answers from the integrated probabilistic graphs.

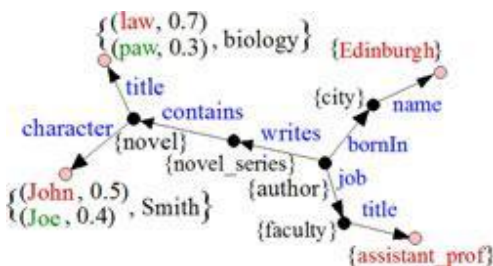


Fig 2: Probabilistic RDF graph:

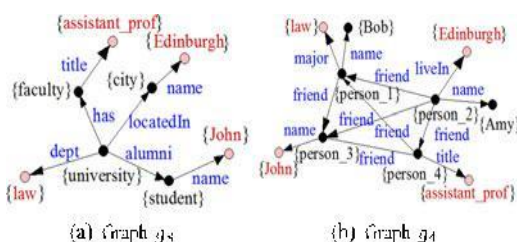


Fig 3: RDF data graphs g_3 and g_4 (keywords: “John”, “law”, “Edinburgh”, and “assistant_prof”).

In this paper, we make the following contributions:

- 1) We enhance the pg-KWS problem in probabilistic RDF graphs with appropriate ranks.
- 2) We design effective pruning strategies to reduce the pg-KWS search space.
- 3) We propose an efficient approach to process pg-KWS queries via pre-computed data.
- 4) We conduct extensive experiments to show the pg-KWS query performance.

II. RELATED WORK

2.1 Data Model for Probabilistic RDF Graphs

A probabilistic RDF graph database G consists of probabilistic RDF graphs G , which are modeled by Bayesian networks. V is a set of vertices v_i , each associated with a set of possible keywords (denoted as $k(v_i)$). In this paper, we consider uncertain keywords in vertices only, and certain labels in edges. In the case where edge labels are uncertain (denoted by CPTs), we can easily extend our solution to the one on a transformed probabilistic RDF graph. That is, we can add a new (dummy) vertex to each edge in probabilistic RDF graph, and assign CPTs of uncertain edge keywords to this new vertex. This way, we can transform probabilistic RDF graph with uncertain vertex/edge keywords to the one with uncertain keywords in vertices only, to which we apply our proposed approaches.

2.2 Probabilistic RDF Keyword Search Queries

The r-Radius Graph Recently, EASE [25] has been proposed to answer keyword search queries on certain graphs. Keyword search answers in EASE are related to the notion of r-radius graph.

2.3 Ranking Scores for RDF Keyword Search

Consider two scores, feature δg_c and entropy δg_e , in a certain subgraph g_c below. Later, we will generalize it to probabilistic RDF graph.

Feature score:

The first dimension, feature δg_c , in the 2D ranking score vector considers both structural and IR features of subgraph g .

2.4 The pg-KWS Problem

The r-radius graph [25] is the basic subgraph unit that we want to obtain for the keyword search in RDF graphs. The r-radius graph centers at some vertex in the graph and has the breadth first traversal depth (radius) not greater than r . Specific to our problem where the RDF graph is probabilistic, we need to consider the ranking score under the possible worlds semantics [9].

Given a probabilistic RDF graph G , a set of q query keywords $k_1, k_2, \dots, k_q$, and a probabilistic threshold $\alpha \in [0, 1]$, a keyword search query over probabilistic RDF graph G (pg-KWS) retrieves r-radius graphs $g \subseteq G$ such that g contains all the query keywords, and h feature δg_c ; entropy δg_e is not

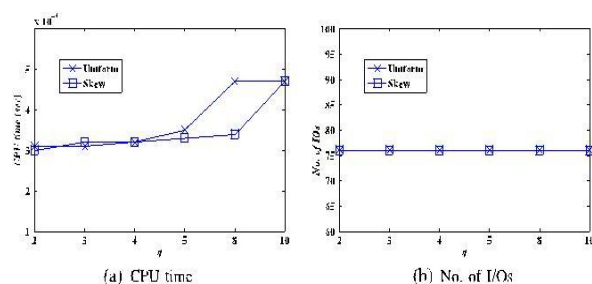


Fig 6: The pg-KWS performance versus q.

In probabilistic/ uncertain databases, the data uncertainty can be classified into two categories [39], attribute uncertainty and tuple uncertainty. Uncertain databases [8] consider the attribute uncertainty, in which uncertain objects are modeled by uncertainty regions. The tuple uncertainty is captured by probabilistic databases that contain relational tuples associated with existence probabilities. Keyword search over graphs has recently been very hot due to its ease of use by users in real applications [6], [16], [22], [25]. Basically, these works return connected keywords in a graph with high ranks, and aim to improve the query efficiency. Query answering on probabilistic/uncertain data considers the possible worlds semantics [9]. That is, a possible world is a materialized instance of the probabilistic database that may appear in the real world. A probabilistic query is equivalent to first answering such a query in all possible worlds individually, and then combining (aggregating) query answers returned from these possible worlds.

V CONCLUSION

We formulate and deal with an important prob of keyword search over probabilistic RDF graphs, called pg-KWS queries. We propose effective pruning methods via offline pre-computed score bounds and probabilistic threshold to quickly filter out false alarms. Moreover, we construct an index for the precomputed data for RDF, and present an efficient query answering approach. Experiments have been conducted to verify the proposed approaches.

REFERENCES

- [1] W3C: Resource description framework (RDF). [Online]. Available: <http://www.w3.org/RDF/>, Feb. 2014.
- [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] 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.
- [4] V. Bicer, T. Tran, and R. Nedkov, "Ranking support for keyword search on structured data using relevance models," in Proc. 20th ACM Conf. Inform. Knowl. Manage., 2011, pp. 1669–1678.

- [5] S. Borzsönyyi, D. Kossmann, and K. Stocker, "The skyline operator," in Proc. 17th Int. Conf. Data Eng., 2001, pp. 421–430.
- [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] N. N. Dalvi and D. Suciu, "Efficient query evaluation on probabilistic databases," The Int. J. Very Large Data Bases, vol. 16, no. 4.
- [8] A. Polleres, "From SPARQL to Rules (and Back)," Proc. 16th Int'l Conf. World Wide Web (WWW), 2007.
- [9] B.B. Dalvi, M. Kshirsagar, and S. Sudarshan, "Keyword Search on External Memory Data Graphs," Proc. VLDB Endowment, vol. 1, pp. 1189–1204, 2008.
- [10] S. Duan, A. Kementsietsidis, and K. Srinivas, O. Udrea, "Apples and Oranges: A Comparison of RDF Benchmarks and Real RDF Datasets," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2011.
- [11] H. Fu and K. Anyanwu, "Effectively Interpreting Keyword Queries on RDF Databases with a Rear View," Proc. 10th Int'l Conf. Semantic Web (ISWC), 2011.
- [12] K. Golenberg, B. Kimelfeld, and Y. Sagiv, "Keyword Proximity Search in Complex Data Graphs," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2008.
- [13] Y. Guo, Z. Pan, and J. Heflin, "LUBM: A benchmark for OWL Knowledge Base Systems," J. Web Semantics, vol. 3, pp. 158–182, 2005.
- [14] H. He, H. Wang, J. Yang, and P.S. Yu, "Blinks: Ranked Keyword Searches on Graphs," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '07), 2007.
- [15] 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), 2003.
- [16] P. Jeavons, "On the Algebraic Structure of Combinatorial Problems," Theoretical Computer Science, vol. 200, pp. 185–204, 1998.
- [17] M. Kargar and A. An, "Keyword Search in Graphs: Finding R-cliques," Proc. VLDB Endowment, vol. 4, pp. 681–692, 2011.
- [18] G. Li, B.C. Ooi, J. Feng, J. Wang, and L. Zhou, "Ease: Efficient and Adaptive Keyword Search on Unstructured, Semi-structured and Structured Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2008.
- [19] Y. Luo, W. Wang, and X. Lin, "SPARK: A Keyword Search Engine on Relational Databases," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), 2008.
- [20] Y. Tao, R. Cheng, X. Xiao, W. K. Ngai, B. K., and S. Prabhakar, "Indexing multi-dimensional uncertain data with arbitrary probability density functions," in Proc. 31st Int. Conf. Very Large Data Bases, 2005, pp. 922–933.
- [21] T. Tran, H. Wang, S. Rudolph, and P. Cimiano, "Top-k exploration of query candidates for efficient keyword search on graph haped (RDF) data," in Proc. IEEE Int. Conf. Data Eng., 2009, pp. 405–416.
- [22] O. Udrea, V. S. Subrahmanian, and Z. Majkic,

- “Probabilistic RDF,” in Proc. Int. Conf. Info. Reuse Integ., 2006, pp. 172–177.
- [24] 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.
- [25] 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.

