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. 3, Issue 7, July 2015

Searching Keyword on Uncertain Graph Data by Using Enhanced Approach

Ashwini V. Urade[1], Pravin Kulurkar[2]

M. Tech Student, Department of Computer Science and Engineering, VIT, Nagpur, India[1]

Assistant Professor, Department of Computer Science and Engineering, VIT, Nagpur, India[2]

ABSTRACT: In various search mechanism keyword search is used and provides a simple but user friendly interface to extract or retrieve information from complicated data structure. As datasets are represented by trees and graph, but in real life application the graph of that datasets are not certain. It is subjected to uncertainties due to incompleteness and ambiguity of data. Because of its uncertainty, it is difficult task to retrieve keyword on uncertain graph, also it provides unwanted result. To overcome from this failure or drawback, this paper used new techniques. This technique provides effective result for searching keyword on graph. Uncertain graph is used in PPI network, modeling Road network, RDF data and social network etc. This technique takes less processing time and search the keyword with efficiency as compared to previous research. Approximate mining algorithms i.e. K-Medoids Algorithm can be used to form sub graph from uncertain graph data based on scores at the level of keywords, data elements, element sets, and sub graphs that connect these elements. To retrieve the efficient keyword from sub graph keyword matching algorithm i.e. selection sampling can be used for uncertain graph data. The objective of propose technique is to reduce the high cost of processing keyword search queries on uncertain graph data and improve the performance of keyword search, without compromising its result quality. Also o reduce processing time for keyword search in uncertain graph data.

KEYWORDS: Database, algorithm, uncertain data, graph data, Keyword searching.

I. INTRODUCTION

As very large amount of data is available from different information sources such as the social media, web, communication networks, software repositories, citation and collaboration networks, there is essential need to query and analyze such data. Much of the data in these domains expresses more complex relationships between objects, making it natural to model it as “graphs". Such data is often noisy and incomplete due to different reasons like due to Missing information or errors from the source, Data extraction errors, Data duplication errors, Data integration errors.

In recent years, the study of keyword search technology based on Graph data has done, and it is generally applied to the field of information retrieval data on worldwide web. In the field of traditional graph database, the research on keyword search has already gained some achievement, but in the field of uncertain graph data, the study on keyword search has just started. However, all graphs in the database are assumed to be certain or accurate, and in real-life applications, this assumption is often invalid. For example, RDF data can be highly unreliably due to errors in the web data or data expiration. The data may have duplicate information, i.e., sets of nodes that refer to the same real world entity, while queries over such uncertain data require reasoning at the real-world entity semantics. Therefore, it is useful to express and encode different types of uncertainty in a probabilistic model, and also perform soft querying over such uncertain graphs and taking into consideration that multiple nodes may just be references to the same entity.

In the application of the data integration, it is needed to incorporate such RDF data from various data sources into an integrated database. In this case, uncertainties/inconsistencies often exist. Like In social networks, each link between any two persons is often associated with a probability that represents the uncertainty of the link or the strength of influence a person has over another person in viral marketing. In XML data (a tree or graph structure), uncertainties are incorporated in XML documents known as probabilistic XML document (p-document). Keyword searching in RDF data, social networks and XML data has many important applications.

Therefore, it is necessary to relax the strict assumption of Deterministic or well certain graphs and study keyword search over uncertain graphs. Keyword Query Analysis and Mining sub-graph pattern is the ultimate goal of research

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. 3, Issue 7, July 2015

on uncertain graph data management to retrieve the useful data from uncertain graph data. The keyword routing method can be used to route keywords only to relevant sources to reduce the high cost of processing keyword search queries over all sources.

A keyword-element relationship summary that compactly represents relationships between keywords and the data elements mentioning them. A multilevel scoring mechanism can be used 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.

To overcome these issues, we propose a new technique for searching keyword on uncertain graph data. For this we use mining algorithm for creating sub-graph from uncertain graph.

II. RELATED WORK

In literature, we study most of the recent mining andsampling techniques that have been developed in data mining domain.

The work in [1] presented efficient keyword searching over uncertain graph data using filtering and verification methods. Where filtering includes three pruning phases existence, path-based and tree-based probabilistic pruning phases. And for verification, the sampling algorithm is used. In existence probabilistic pruning, all uncertain information is removed from the graph.

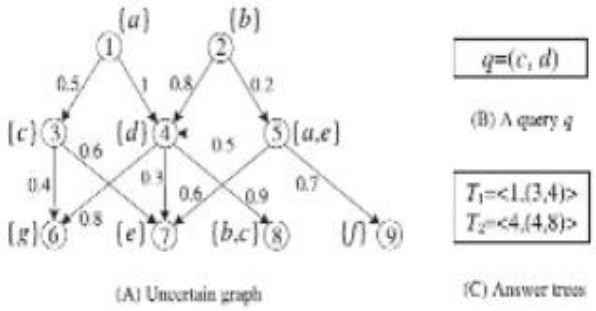

Fig.1. Example of query and answers.

Fig. 1A shows an uncertain graph g with each node attached some text, i.e. node 8 containing two keywords {b,c}. A real number associated with each edge represents the existence probability of the edge, i.e. 0.5 denoting the existence probability of edge (1,3) Fig. 1B shows a keyword query, q = (c,d).

O. Papapetrou in [2] presenteted A method that uses an index of the uncertain graph database to reduce the number of comparisons needed to find frequent subgraph patterns. The proposed algorithm relies on the apriori property for enumerating candidate subgraph patterns efficiently. Then, the index is used to reduce the number of comparisons required for computing the expected support of each candidate pattern. It also enables additional optimizations with respect to scheduling and early termination, that further increase the efficiency of the method.

Lei Zhang in [3] presented A novel method for computing top-k routing plans based on their potentials to contain results for a given keyword query. Also propose to route keywords only to relevant sources to reduce the high cost of processing keyword search queries over all sources. 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 that connect data elements.

Zhao Zhibin in [4] presented An optimized algorithm DMPUTop-k for processing most probable uncertain Top-k queries in the distributed environment.

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. 3, Issue 7, July 2015

J. Li and S. Zhang in [5] introduced an efficient approximation algorithm to determine whether a sub-graph pattern can be output or not. This paper propose also propose efficient sampling algorithm for matching the keyword in sub-graph.

Jun Gao [6] introduced a FEM framework to bridge over the gap between graph operations and relational operations. To improve the performance of FEM framework, new feature of SQL standards viz. window function and merge statements are introduced in this paper. An edge weight aware graph partitioning schema and design a bi-directional restrictive BFS (breadth-first-search) over partitioned tables, are proposed which helps to improve the scalability and performance by avoiding extra indexing overheads. Likewise, keyword searching over uncertain graph data becomes easier.

All these techniques tried to cover different issues maintaining the cost of implementation but it requires more time and the high cost of processing keyword search queries on uncertain graph data.

III. PROPOSED ALGORITHM

The objective of proposed techniques is to search keyword over uncertain graph data and reduce the high cost of processing keyword search queries on uncertain graph data.

This project has divided in four modules, file processing, generation of uncertain graph, formation of sub-graph and finally search keyword over graph consisting tree.

1. File Processing

In proposed System We are going to form an uncertain graph of an uncertain Data and search keyword over an graph, The First step of it is to process a file and search the keyword over a file content by sentence wise and paragraph wise.The searching result is saved in a datacenter as an uncertain data, and for creation of uncertain data, data is collected in database from text file, pdf, document file etc. In this module we process data and store in database. This data is further used for creation of uncertain graph. The data with ambiguity stored in database.

2. Generation of Uncertain graph

Graph creation can be performed on stored data in database. This data arranged in graph or tree form by using tree based approach in which parent node have its child. All keywords is plotted below its parent node, here the parent node is uncertain graph data. After preprocessing the complete data, the data is clustered and by clustering, it becomes easy to generate a graph. All keywords is plotted below its parent node, here the parent node is uncertain graph data.

3. Formation of sub-graph

Sub-graph can be plotted by removing ambiguity in uncertain graph. This can be done with the help of K-mediods algorithm which find out mediods and plot sub-graph. This algorithm also use to reduce ambiguity.

3.1 K-Medoids Algorithm

The *k*-mediod algorithm is a clustering algorithm related to the *k*-means algorithm and the medoid shift algorithm. It minimizes a sum of general pair wise dissimilarities instead of a sum of squared Euclidean distances.

By using this technique we finding mediods and go for clustering the overall data. In proposed System we get the occurrences of each and every word in our uncertain dataset and try to get pruning of data to get simplest subgraph. The subgraph levels are assigned and design over the medoids and cluster for formation of graph.

The most common realisation of *k*-medoid clustering is the **Partitioning Around Medoids (PAM).**

1. Initialize: randomly select *k* of the *n* data points as the medoids**.**

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. 3, Issue 7, July 2015

2. Assignment step: Associate each data point to the closest medoid. Closest defined using any valid distance metric like Manhattan distance.

3. Update step: For each medoid *m* and each data point *o* associated to *m* swap *m* and *o* and compute the total cost of the configuration (that is, the average dissimilarity of *o* to all the data points associated to *m*). Select the medoid *o* with the lowest cost of the configuration.

Repeat alternating steps 2 and 3 until there is no change in the assignments.

Where cost between any two points is found using formula

$$\text{cost}(x, c) = \sum_{i=1}^{d} |x_i - c_i|$$

Where *x* is any data object, *c* is the medoid, and *d* is the dimension of the object which in this case is 2.Total cost is the summation of the cost of data object from its medoid in its cluster.

4. Keyword searching on sub-graph

In proposed system we are going to Sample i.e. Refined the Uncertain Data which is present in a dataset. We use the Selection Sampling in this proposed system.

4.1. Selection sampling

In this technique we Classify data into having two major class of "Having Class (Keyword Path Found)" and "Not Having Class (Keyword Path Not Found) Class" behaviour. The data is refined over majority class of having, which gives the path of keyword for searching over an uncertain Data. The new data path of that Keyword is got through the Sampling approach.

1. Get the dataset to inputs.
2. Classify data into having two major class of having Class and not having Class behavior.
3. The data is refined over majority class of having, gives the path of keyword for searching over an uncertain Data.

In the given approach as per the input query keywords, the algorithm scan the entire graph from root node to leaf nodes till reaching to the all keyword. It maintains an index to store all the keywords in database and finally shows the sub tree in output result.

IV. SIMULATION RESULT

The simulation result shows the working of proposed system for searching keyword on uncertain graph data by using a proposed algorithm. In this approach mining and sampling algorithm is used for searching keyword on uncertain graph data.

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. 3, Issue 7, July 2015

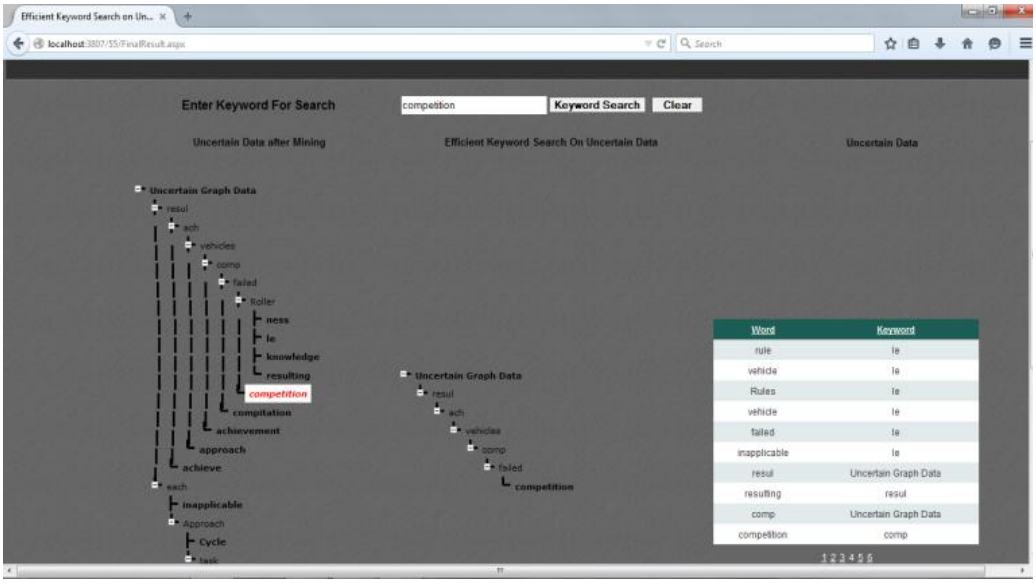

Fig. 2 Resulting keyword with its sub tree

The above Fig.2 shows the experimental result of proposed system, in which it searches keyword from uncertain graph data and shows the result with its sub tree. The keyword is search by using sampling technique and resultant graph is generated by using.

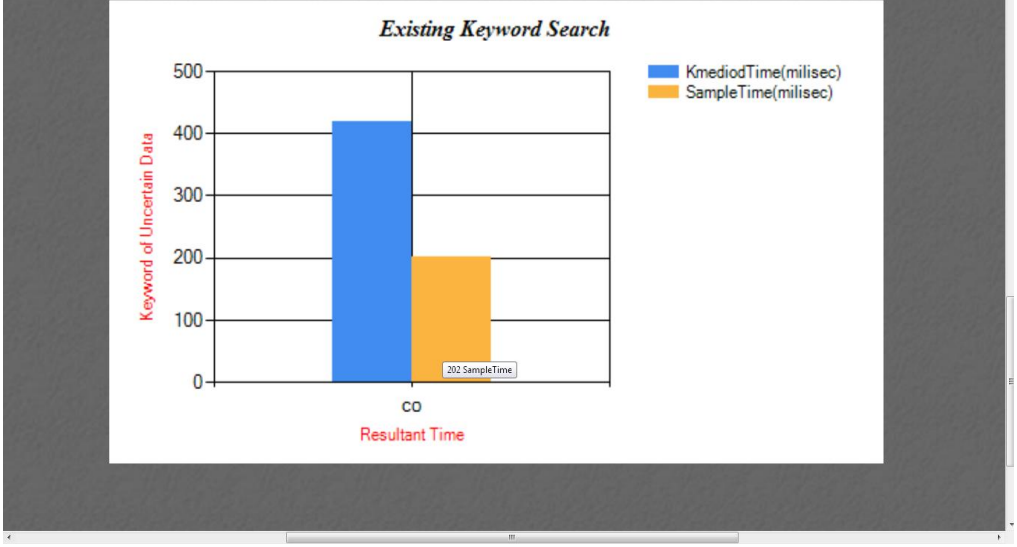

Fig 3. Time taken for Existing keyword Search

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. 3, Issue 7, July 2015

In above analysis graph fig 3, the results are shown about time taken for searching keyword on uncertain graph in which uncertain data is already available in dataset. As the Keyword is query is encounter it takes K-Medoid and sampling time for searching keyword. The time required for keyword searching is in millisecond, it means the more data and matching data is found out at high percentage within less time with this approach.

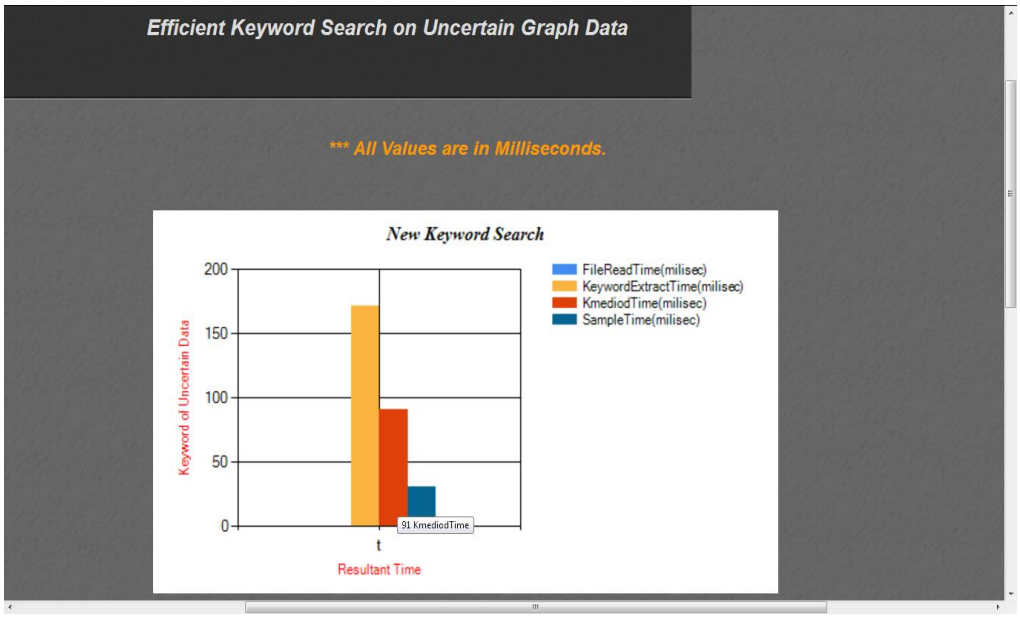

Fig 4. Time taken for New Keyword search

In above analysis graph (fig 4.), the results are shown about time taken for overall working of propose approach. In which dataset is creates at runtime by file processing and uncertain graph is generated. In this subgraph is form by K-medoid and keyword is search by sampling technique. As the Keyword is query is encounter it takes K-Medoid and sampling time for searching keyword. The time required for these processing is in millisecond, it means the more data and matching data is found out at high percentage within less time with this approach.

V. CONCLUSION AND FUTURE WORK.

The simulation results showed that the proposed algorithm performs better give the efficient result including keyword and its sub tree. Many real life datasets are represented by trees and graphs, keyword search has become an attractive mechanism for data of a variety of types. Because of the underlying graph structure, keyword search over graph data is much more complex than keyword search over documents. So proposed work is about searching keyword in an uncertain graph data with preprocessed keyword query. The keywords are searched on graph and generate the subtree which includes all keywords. The proposed approach provides the efficient results for user keyword query.

An approach has presented in this project which is used to derive the relevant data from graphs by keyword searching. As many real life applications are there in graph like road networks, social media networks and PPI networks, so the proposed approach can be used on these real life applications to search any element on huge graph and can get the reasonable and relevant sub graph in results so it will also be helpful to retrieve the subgraphs from these complex and real graphs. As in road networks it can find out the shortest distance route between source to destination by retrieving the exact sub graph by applying the proposed algorithm on it with the manipulations of distances.

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. 3, Issue 7, July 2015

REFERENCES

1. Ye Yuan, Guoren Wang, Lei Chen, and Haixun Wang," Efficient Keyword Search on Uncertain Graph Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013.
2. O. Papapetrou, E. Ioannou, and D. Skoutas, "Efficient Discovery of Frequent Sub-graph patterns in Uncertain Graph Database" Proc. 14th Int'l conf. Extending Database Technology (EDBT), 2011.
3. Thanh Tran And Lei Zhang," Keyword Query Routing", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 2, February 2014.
4. Zhao Zhibin, Yu Yang, Bao Yubin, Yu Ge," Optimizing Distributed Top-k Queries on Uncertain Data", IEEE,2013.
5. Z. Zou, H. Gao, J. Li, and S. Zhang, "Mining Frequent Subgraph Patterns from Uncertain Graph Data," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 9, pp. 1203-1218, Sept. 2010.
6. Jun Gao, Jiashuai Zhou, Jeffrey Xu Yu, and Tengjiao Wang,"Shortest Path Computing in Relational DBMSs", IEEE vol. 26, no. 4, April 2014.
7. K. Yi, F. Li, D. Srivastava, and G. Kollios, "Efficient Processing of Top-K Queries in Uncertain Databases," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), 2008.
8. G. Kollios, M. Potamias, and E. Terzi, "Clustering Large Probabilistic Graphs," IEEE Trans. Knowledge and Data Eng ,Feb. 2013.
9. E. Adar and C. Re, "Managing Uncertainty in Social Networks," IEEE Data Eng. Bull., vol. 30, no. 2, pp. 15-22, June 2007.
10. M. Potamias, F. Bonchi, A. Gionis, and G. Kollios, "K-Nearest Neighbours' in Uncertain Graph," Proc. VLDB Endowment, vol. 3, pp. 997-1008,2010.
11. Mayssam Sayyadian, Hieu LeKhac, AnHai Doan, Luis Gravano," Efficient Keyword Search Across Heterogeneous Relational Databases", IEEE 2007.
12. Lifang Qiao, Yu Wang,"A Keyword Query Method for Uncertain Database", 2nd International Conference on Computer Science and Network Technology, IEEE, 2012.
13. Bolin Ding, Jeffrey Xu Yu, Shan Wang, Lu Qin, Xiao Zhang, Xuemin Lin," Finding Top-k Min-Cost Connected Trees in Databases", IEEE 1-4244-0803-2/07/2007.
14. Branimir T. Todorovic, Svetozar R. Rancic, Ivica M. Markovic, Eden H. Mulalic, Velimir M. Ilic, "Named Entity Recognition and Classification using Context Hidden Markov Model," 9th Symposium on Neural Network Application in Electrical Engineering, NEUREL, pp. 43-46, 2008