KEYWORD BASED SEARCHING ON UNCERTAIN GRAPH DATA BY ENHANCED APPROACH

Miss. Ashwini Urade[1], Prof.Pravin Kulurkar[2]

[1]*Mtech 4th sem, CSE,Vidarbha Institute of Technology, Maharashtra, India,* ***rupaliurade88@email.com***
[2]*Asst professor, CSE,Vidarbha Institute of Technology, Maharashtra, India,* ***pravinkulurkar@email.com***

Abstract

As in various search mechanism keyword search is used which provides a simple but user friendly interface to retrieve information from complicated data structure. Since many datasets are represented by trees and graph, but in real life application this graph are not certain. It is subjected to uncertainties due to incompleteness of data. Because of its uncertainty, it is difficult task to search keyword on uncertain graph, also it provides unwanted result. To overcome from this 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 as compared to previous research.
Approximate mining algorithms 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 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.*

---------- *** ----------

1. INTRODUCTION

As large amount of data is available from different information sources such as the Web, social media, 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 considerationthat 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 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.

2. RELATED WORK

In literature, we study most of the recent mining and Sampling techniques that have been developed in data mining domain. The filtering-and-verification framework to answer the query on uncertain graph data.[1].

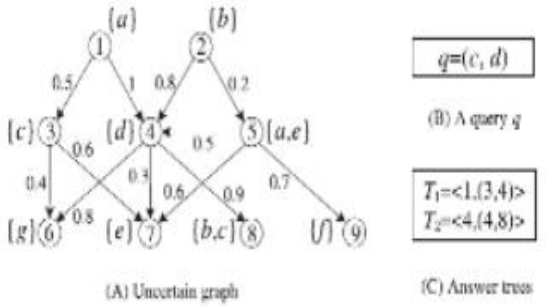

Fig-1. Example of query and answers.

A method that uses an index of the uncertain graph database to reduce the number of comparisons needed to find frequent sub-graph patterns by using Apriori property. [2].A novel method for computing top-k routing plans based on their potentials to contain results for a given keyword query[3]. An optimized algorithm DMPU Top-k for processing most probable uncertain Top-k queries in the distributed environment. [4]. Anefficient approximation algorithm to determine whether a sub-graph pattern can be output or not. [5].

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

3. PROBLEM DEFINITION

There is no work on keyword search in uncertain graph data. For keyword searching in uncertain graph database, two phases were used which are filtering and verification. For filtering purpose, there were also sub phases which are existence probabilistic prune, path based probabilistic prune and tee based probabilistic phase which consumed more time for filtering and finally verification is applied. This procedure consumed much more time.

4. PROPOSED WORK

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.Also to improve the performance of keyword search, without compromising its result quality and to reduce processing time for keyword search in uncertain graph data.

To achieve the objective of this project, we have proposed following techniques;
-While creating the sub graph over uncertain graph data, we are going to find an approximate set of frequent sub graph patterns in graph database by using K-Medoids algorithms.
-Then we find the efficient keyword in this sub-graph by using Efficient selection sampling technique. This algorithm is used to verify the candidate.

5. WORK DONE

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

5.1. File Processing

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.

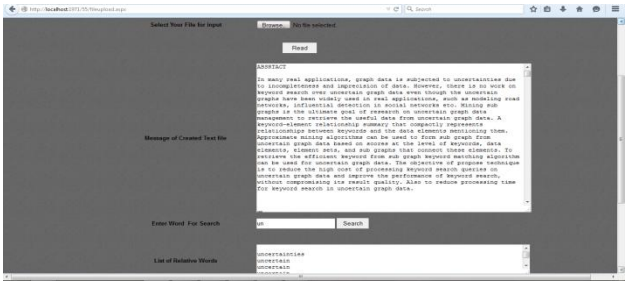

Fig-2. Snapshot of File Processing

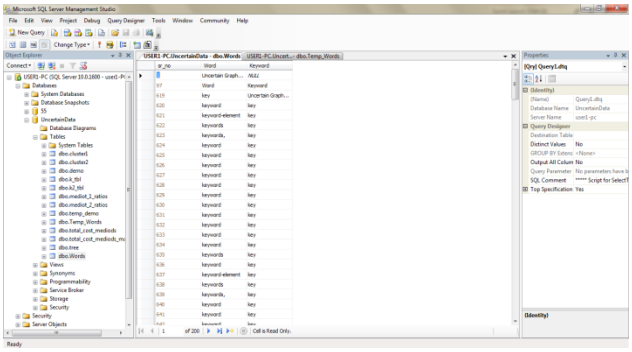

Fig-3.Snapshot of database of uncertain data

5.2. Creation of Uncertain graph

Creation of graph 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.

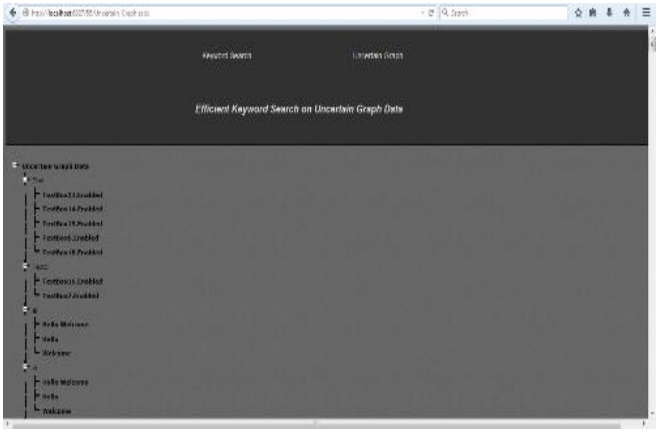

Fig-4. Snapshot of uncertain graph

5.3. 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.

5.3.1 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.

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**.**

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.

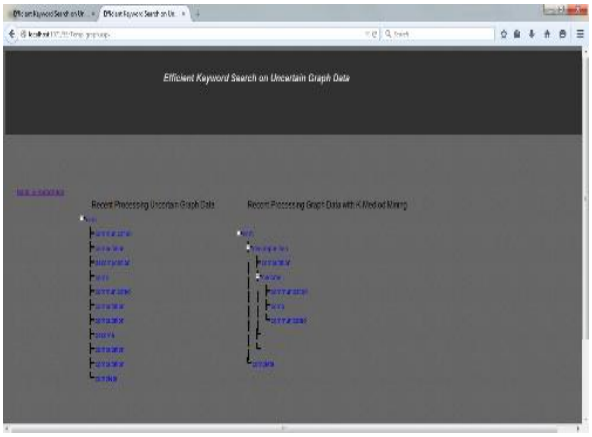

Fig-5. Snapshot of Sub-Graph

5.4. Keyword searching on sub-graph

Keyword searching can be performed over sub-graph by selection sampling technique. This technique searches the query keyword over sub-graph, if it found then it shows the path tree of that keyword as an output.

6. CONCLUSION

This paper proposes a technique to perform keyword based search on uncertain graph using mining and sampling algorithms. We provide K-Medoid algorithm for sub-graph formation and selection sampling technique for matching keyword over sub-graph. It also improve the performance of keyword search, without compromising its result quality, also reduce processing time.

7.REFERENCE

[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'1 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, BaoYubin, 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] 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.

[7] G. Kollios, M. Potamias, and E. Terzi, "Clustering Large Probabilistic Graphs," IEEE Trans. Knowledge and Data Engg, Feb. 2013.

[8] E. Adar and C. Re, "Managing Uncertainty in Social Networks," IEEE Data Eng. Bull., vol. 30, no. 2, pp. 15-22, June 2007

[9] M. Potamias, F. Bonchi, A. Gionis, and G. Kollios,"K-Nearest Neighbours' in UncertainGraph, " Proc. VLDB Endowment, vol. 3, pp. 997-1008,2010

[10] MayssamSayyadian, HieuLeKhac, AnHai Doan, Luis Gravano," Efficient Keyword Search Across Heterogeneous Relational Databases", IEEE 2007.

[11]H. He, H. Wang, J. Yang, and P.S. Yu, "Blinks: Ranked Keyword Searches on Graph" Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 305-316, 2007.

[12] B. B. Dalvi, M.Kshirsagar, and S. Sudarshan, "Keyword Search on External Memory Data Graph," Poc. VLDB Endowment, vol. 1, pp. 1189-1204, 2008.

[13]B.K.K. Golenberg and Y. Sagiv, "Keyword Proximity Search in Complex Data Graph," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2008.

AshwiniUradereceived the B.E. degree from KDKCE, RTM Nagpur University.Maharashtra, India. Sheis currently working toward the Master's degree with the department of Computer Science and Engineering in Vidarbha Institute of Technology, RTM Nagpur University, Maharashtra, India.
Her research interests include Data mining,Web mining, Graph mining

Prof. Pravin Kulurkar received BE degree And Master's degree from RTM Nagpur University, Maharashtra, India.
He is currently working with department of Computer science and Engineering as an Assistant professor for Vidarbha Institute of Technology, RTM Nagpur University, Maharashtra, India.