e-ISSN: 2349-9745
p-ISSN: 2393-8161
Scientific Journal Impact Factor (SJIF): 1.711
International Journal of Modern Trends in Engineering and Research
www.ijmter.com

INDEXING ON UNCERTAN DATA USING U-QUADTREE

Mr. Sangram Patil[1], Mrs. R. J. Deshmukh[2]
[1]*Department of Technology, Shivaji University, Kolhapur*
[2]*Department of Technology, Shivaji University, Kolhapur*

Abstract: In recent year, in many applications like radio frequency identification (RFID) networks and location-based services (LBS), it is highly demanded to address the uncertainty of the objects. For effective indexing of such uncertain one dimensional objects MV-Tree, US+ Tree [2] structure can be used which helps in effectively searching of them. These structures are used for Range Query, Nearest Neighbor Query. But these structures are not used if uncertain objects are in multidimensional structure. More focus is on indexing multidimensional uncertain objects for effective and efficient searching. So for this using a novel indexing structure, named U-Quadtree [1], for efficient processing of uncertain object in multidimensional space novel indexing through U-Quadtree is prepared. In this a cost model which carefully considers various factors that may impact the performance. An effective and efficient index construction algorithm is proposed to build the optimal U-Quadtree regarding the cost model. U-Quadtree can also efficiently support other types of queries such as uncertain range query and nearest neighbor query.
Keywords – Uncertain Data, U-Quadtree, Range Query, Uncertain Range Query, KNN Search.

I. INTRODUCTION

In recent year due to fast growth in the technology, computer plays an important role in every business. In much business, data plays an important role for further enhancement of the business. As number of years passes by, data goes on increasing in size. To store this large data many advanced techniques been developed. The required data can be collected personally or appointing agency/person or through any sensing device which indicates there are multiple ways of collecting the data. In many application data can be collected by sensing devices / radio frequency devices. Due to error in sensing device, due to poor quality of sensing data, due to overlap of region of sensing device, data collected contain error or partially incomplete which is called as uncertain in nature. It might be possible that data get error or lost in transmission system. This has created a need for uncertain data management system or applications [12]. Uncertain data is ubiquitous in many real world applications, such as environment monitoring, transportation system, sensor network, market analysis, medical diagnosis etc. These all application generates data at an alarming rate [5]. In the era of explosive growth of information, how we store the massive data, and we access the information is most important part for user / application. We can use any data base management system to store the data. But the main problem is not to store the data but to effectively fast retrieval of the data which can be achieved partially by using fast storage device. As database grows query takes more time for execution. One of the most effective and ubiquitous tool for reducing query execution time in traditional database system is indexing. Virtually all database systems support variety of index structures like B-Tree, hash indexing, R-Trees [2]. Each index structure can be used for certain types of data and query types. For uncertain data we have to design effective indexing structure for storing and retrieving of the data.

There are many types of queries which we can execute on uncertain data like point query, range query, top-k query and probabilistic nearest neighbors query. In many applications like location based services (LBS), global position system (GPS), sensor data analysis range searching problem is fundamental problem. For effective and accurate search new index technique is used called U-Quadtree. This index structure is based on the quadtree because it is flexible data structure

in the sense that can adaptively build summaries of the objects so that overall cost of the range searching can be minimized regarding the cost model using U-Quadtree [1]. This index structure will help to effectively use of range query and nearest neighbor search on uncertain objects.

II. NEED FOR INDEXING

Radio frequency identification (RFID) networks or sensor networks are widely used in many places/ applications. The range searching problem is fundamental in these types of applications. In many location based system the objects and their current location can be obtained by sensor devices or some type of readers. There are many factors like low cost readers or some nearby metal object can predict uncertain region of the object which might give the wrong result when queried. Exiting work follow the filtering and verification paradigm which evolves filtering cost in terms of I/O or CPU cost. Considering this cost an effective index structure must be used for effectively indexing the objects in the database to get fast and accurate response for range query for uncertain objects. This indexing must be in such a way that range query and nearest neighbor query must work effectively and accurately [1].

III. SYSTEM ARCHITECTURE

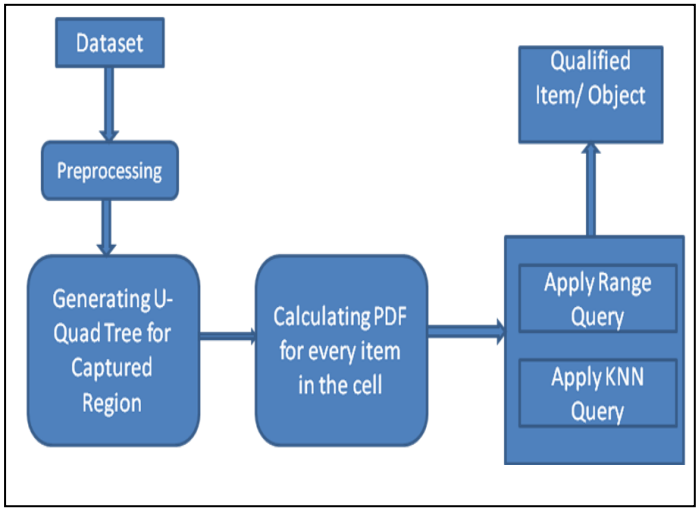

Figure 1: System Architecture

A Quadtree is a tree data structure in which each internal node has exactly four children. Quadtree are most often used to partition a two-dimensional space by recursively subdividing it into four quadrants or regions. The regions may be square or rectangular, or may have arbitrary shapes. A novel index structure, called U-Quadtree, can be used to effectively organize the multidimensional uncertain objects with arbitrary PDFs. Designing a cost model to quantitatively analyze the performance of the range search, followed by efficient optimal U-Quadtree construction algorithm.

U-Quadtree can also support uncertain range search where the range search region is uncertain as well as the k nearest neighbor search on uncertain objects. Comprehensive experiments demonstrate the efficiency of the U-Quadtree technique. U-Quadtree can also efficiently support other types of queries such as uncertain range query and nearest neighbor query.

The empirical study shows that existing index techniques can significantly improve the performance of the range search. Nevertheless, the “equality strategy” is employed in the existing techniques; that is, the same amount of resources in terms of this cannot address the problem of “equality strategy” since all objects still have the same summary size. Moreover, it is nontrivial to adapt the existing indexing techniques to the biased resource allocation strategy for uncertain objects[1]. Therefore, it is desirable to develop new indexing technique for uncertain objects such

that the summaries of the objects can be constructed in a flexible way. Intuitively, the more resources (i.e., larger summary size in terms of space usage) assigned to the summary of an uncertain object U, the tighter the lower and upper appearance probability bounds of U and hence the lower expected verification cost of U[2]. On the other hand, the filtering cost, which is usually dominated by the index I/O delay, increases with the storage overhead of the summaries. Therefore, the key of the index construction is to find a good tradeoff between filtering cost and verification cost such that the overall cost is minimized.

Quad Tree

A quad tree is a space partitioning tree data structure in which a d-dimensional space is recursively subdivided into 2d regions (cells). Due to its simplicity and regularity, the quad tree technique has been widely applied in many applications. In the paper, we focus on two-dimensional space and all techniques developed can be immediately applied to higher dimensional spaces. In our work, we say a U-Quadtree [1] is optimal regarding the cost model if is minimized. In applying using quad tree, it can be investigate how to efficiently construct the optimal U Quadtree regarding the cost model so that the overall cost of the range search can be minimized.

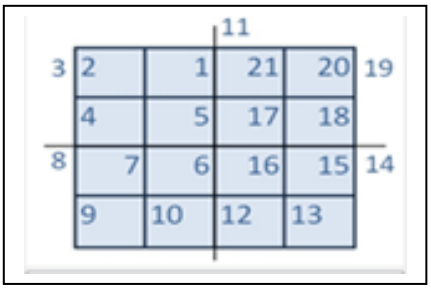

Figure 2: U-Quadtree Structure Numbering

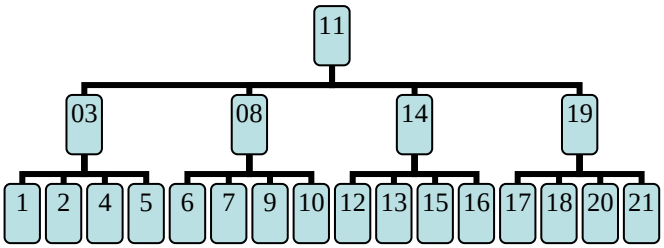

Figure 3 : U-Quadtree Structure

Entry Index:

A tree used to keep entries of the objects in the secondary memory, where the key of each entry is its cell id. Similar to, we assume the id of a cell is its Hilbert code[3] generated in a recursive way such that the cells with close spatial proximity are likely to be allocated to the same or adjacent pages in UQE[1]. Particularly, a leaf node of UQE is called the entry page[1] and f denotes its capacity (i.e., the maximal number of entries in an entry page).

Range Query:

In Range Query[1] the lower and upper bounds must be derived for range search following the filtering-and-verification. Retrieves a set of nonempty cells which are contained or overlapped by range query. This can come up with the lower and upper bounds of the appearance probabilities of the objects and the cost to compute lower and upper bounds of the appearance probabilities.

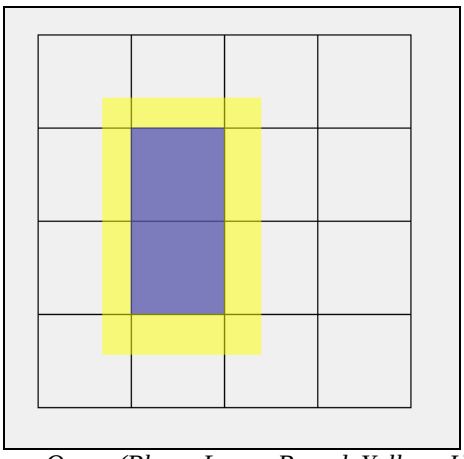

Figure 4: Range Query (Blue – Lower Bound, Yellow- Upper Bound)

Algorithm for Range Query

Input: **UQ:** the UQuad tree set region from U sets

rq: Range for Query

q : Threshold Value

Output: Set of items with Papp(U, rq)>= θ.

Step 1: Get the Region for Range Query

Step 2: Get the Threshold Value for Search

Step 3: Get Lower Bound (LB)

Step 4: Get Upper Bound (UB)

Step 5: Remove the items from UB having threshold < given value

Step 6: Get the items from LB having threshold Greater Than given value

Step 7: Get the items from remaining items of UB having threshold Greater Than given value

Step 8: Return the items.

In above algorithm from the given range in step 3 and 4 lower bound and upper bound is derived which will shorten the search area. Then n step no 5 only items whose PDFs is less than the threshold value are omitted and from lower bound only those items whose PDFs is greater than the threshold value are selected. Then in step 7 remaining items are checked for threshold value. Then all selected items are returned.

Uncertain Range Query:

In some applications the range query may be uncertain. For instance, in the location-based service a query point (e.g., a mobile device) may be represented by an uncertain object Q due to the inaccuracy of the measurement. Then, the uncertain range query is represented by an uncertain object Q, a query distance and a probabilistic threshold; that is, find the uncertain objects whose appearance probabilities.

Algorithm for Uncertain Range Query

Input: **UQ**: the UQuad tree set region from U sets

Q: Query Object

γ: Distance

θ: Threshold Value

Output: Set of items with Papp(U, Q, γ) >= θ.

Step 1: Get the Query Point Object for Query

Step 2: Get the Threshold Value for Search

Step 3: Get the Distance for Search

Step 4: Get the Cells (LB) with the range/distance

Step 5: Get items from above cells present within the range

Step 6: Calculate the PDFs of the items which we get from above step

Step 7: Get the items from Step6 having PDF >= θ

i.e. Papp(U, Q, γ)>= θ

Step 8: Return the qualified items.

In above algorithm instead of considering Range here Object and area is considered around the selected object. In step 4 Lower Bound is defined and items are separated from above cell in step 5. In step 6 PDF is calculated for the cell and then items are selected in step 7 and returned.

Nearest neighbor search:

Here a new KNN algorithm[1] based on the U-Quadtree. The essential idea is to identify the promising objects based on the summaries of the objects, and then verify the candidates by computing their exact expected ranks regarding the query point q. The computational cost is still expensive if simply calculate the lower and upper bounds[1] of the expected ranks for each object based on their summaries. In the following, a new KNN algorithm is considered which incrementally extends the search region from the query point q to prune objects based on their summaries following the filtering-and verification paradigm[4]. Particularly, two important pruning rules, namely dominance-based pruning rule and rank based pruning rule, will be first introduced.

Algorithm for KNN Search

Input: **UQ**: the UQuad tree set region from U sets

Q : Query Point

θ : Threshold Value

k : Number of items

d : distance for range

Output: Set of items with highest expected rank within the range.

Step 1: Get the Query Point Object for Query

Step 2: Get the Threshold Value, distance and value for k for Search

Step 3: Get the Cells (LB) with the range/distance

Step 4: Get the items from above cells present within the range

Step 5: Calculate the PDFs of the items which we get from above step

Step 6: Get the items from Step 6 having

PDF >= θ

Step 7: Get top k items from Step 6 having

PDF >= θ

Step 8: Return the qualified items.

In above algorithm Object is considered as query point. In step 3 cell is defined for search which will reduce the search objects. Then n step 4 objects within the range are selected and in next step 5 items are selected and in step 7 top k items are retuned.

IV. CONCLUSION

To address the uncertainty in various applications, an effective indexing technique is developed to support range search on multidimensional uncertain objects. Initially given range is divided into quad format form which U-Quad tree is generated which contains cellid and entry page which gives fast searching of required items objects.

First range is captured using GPS device which further divided into quad format. Item wise PDFs are calculated for each cell and further U-Quad tree is generated with PDFs. These PDFs are used in range query. For range query lower and upper bound is calculated and also threshold value is taken from user. For fast search initially items from lower bound are compared with threshold value and those who PDF is greater than Threshold value selected. Then items from upper bound having PDF value lower the threshold value is removed and remaining items are verified with threshold value. Because of U-Quad tree searching becomes very fast for uncertain objects.

REFERENCES

[1]. Ying Zhang, Wenjie Zhang, Qianlu Lin, Xuemin Lin, "Effectiveli Indexing the Multidimensional Uncertain Objects", IEEE transactions on knowledge and data engineering , Vol 26, No. 3, March 2014.

[2]. Chih-Wu Chung, Ching-Hung Pan, Chuan-Ming Liu, "An Effective Index for Uncertain Data", 2014, IEEE DOI 10.1109/IS3C.2014.132

[3]. Rui Zhu, Bin Wang, Guoren Wang, "Indexing Uncertain Data for Supporting Range Queries", WAIM 2014, LNCS 8484, pp.72-83, Springer 2014
[4]. Navya E. K., M. Madhan Kumar, "Processing Uncertain Database Using U-Skyline Mechanism", ISSN Vol. 2 , Issue 4, April 2014
[5]. Andreas Zufle, Tobias Emrich, Klaus Schmid, "Representative Clustering of Uncertain Data", KDD August 2014
[6]. Sunil Prabhakar, Rahul Shah, Sarvjeet Singh, "Indexing Uncetain Data", Chapter 10, 2014
[7]. Xiang Lian, Lei Chen, " Trip Planner Over Probabilistic Time-Dependent Road Networks", IEEE transactions on knowledge and data engineering , Vol 26, No. 8, August 2014.
[8]. Miyoung Jang, Min Yoon, Jae-Woo Chang, " A Privacy-awar Query Authentication Index for Database Outsourcing", IEEE 978-4799-3919/2014
[9]. Jianxin Li, Chengfei Liu, Rui Zhou, " Quasi-SLCA Based Keyword Query Processing over Probabilistic XML Data", IEEE transactions on knowledge and data engineering, Vol. 26, No. 4, April 2014
[10]. James Bornholt, Todd Mytkowicz, Kathryn S. McKinley, "A First-Order Type for Uncertain Data", ACM 978-4503-2305-5/14/03, March 2014
[11]. Xiang Lian, Lei Chen, " Trip Planner Over Probabilistic Time-Dependent Road Networks", IEEE transactions on knowledge and data engineering, Vol 26, No. 8, August 2014.
[12]. Kenta Funaki, Teruhisa Hochin, Hiroki Nomiya, "Parallel indexing of large multi-dimensional data", IEEE 978-0-7695-5071-8, 2013
[13]. Peng Lu, Lidan Shou, Kian-Lee Tan, " An Efficient and Compact Indexing Scheme for Large- scale Data Store", IEEE 978-1-4673-4910-9, 2013
[14]. M. Suresh Krishna Reddy, R. Jayasree, "Extending Decision Tree Clasifiers for Uncertain Data", International Journal of Engineering Science and Advanced Technology, Vol. -2, Issue-4, 1030-1034, August 2012.
[15]. Miss Pragati Pandey, Mrs. Minu Choudhary, " Uncertain Data Algorithms and Applications", IJARCSSE, Research Paper, Vol.- 2, Issue – 7 , July 2012
[16]. Haiping Jiang, Xinghua Fan, " Research and Implementation of Index Scheme for Large-scale Dataset", IEEE 978-1-4244-7255-0,2011
[17]. Charu C. Aggarwal, Philip S. Yu, " A survey of Uncertain Data Algorithums and Applications", IEEE, 2011

