An Analysis of Space Query Classifier Indexing for Mining Uncertain Data

M. Kalavathi
Head, Department of Computer Science
Government Arts and Science College
Komarapalayam
Kalavathisakthi@gmail.com

P. Suresh
Head, Department of Computer Science
Salem Sowdeswari College
Salem
sur_bh0071@rediffmail.com

***Abstract*-Data uncertainty develops into an accepted topic in database and data mining area due to the extensive survival of uncertainty. The uncertain data is used in several real applications such as sensor network monitoring, object recognition, Location-Based Services (LBS), and moving object tracking. Due to the intrinsic property of uncertainty, many interesting queries are used for different purposes. Data uncertainty arises clearly and inherently in many applications. The causes of uncertainty in applications comprise data uncertainty, incompleteness, control of measuring equipment, the delay or loss of data updates and privacy preservation. Hence, this article mainly concentrates to solve the above mentioned tolerance problem and also reduces the overhead count.**

Keywords-Data uncertainty, LBS, Sensor Network monitoring, overhead count

I. INTRODUCTION

Data mining is a process designed to examine the large amount of data collected. It is a collection of tools employed to execute the process. Data are gathered from many areas like marketing, health, communication in data mining techniques. Data mining is the removal of unknown analytical information from large records. It also helps to locate the hidden patterns, predictive information to use the specialists with solution outside their expectations. The aim of data mining is to remove knowledge from dataset in human-understandable structures. In computer science, uncertain data are the conception of data with noise that changes from the exact values.

Data uncertainty is classified into existential uncertainty and value uncertainty. Existential uncertainty emerges when it is uncertain with objects or with the existence of a data tuple. A data tuple in a relational database is linked with a probability which symbolizes the confidence of the existence. Probabilistic databases are applied to semi structured data and XML. Value uncertainty appears when a tuple is identified to exist. However, the values are not identified exactly. A data item with value uncertainty is represented using a pdf in a finite and bounded region of values. Uncertainty is collected using pdf that are denoted by sets of sample values. Mining uncertain data is costly because of information.

With many techniques for modeling uncertain data, tuple and attribute level uncertainty models are used. The attribute values for an information tuple are denoted using a probability distribution over different choices and takes the autonomy across tuples. The growth of tuple and attribute-level uncertainty models is because of the facility to explain complex dependencies while outstanding simple to represent the relational systems with intuitive query semantics. A probabilistic database is a summarizing representation for a probability distribution over an exponentially large collection of possible worlds representing a probable deterministic example of the uncertain data.

This paper is organized as follows: Section II discusses reviews on space query classifier indexing on uncertain data, Section III describes the existing classification techniques on uncertain data, Section IV identifies the possible comparison between them, Section V explains the limitations as well as the related work and Section VI concludes the paper, key areas of research is given as to solve the tolerance problem and to reduce the overhead count.

II. LITERATURE REVIEW

Existing spatiotemporal tolerance for CRQ relax a query's accuracy necessities in terms of a maximal acceptable error which offer well-defined query semantics with different source of information uncertainty. Spatiotemporal tolerance position sensing operations contains significant source of energy consumption which cannot be applied to NN classifier based continuous queries [3]. Superseding Nearest Neighbor (SNN) Search as described in [2] has the finest objects as any object outside the SNN core is not better than all the objects. SNN-core is supportive type of results for NN search, where it is important to minimize the number of reported objects. SNN-core is computed by utilizing a predictable multidimensional index but increased the overhead. SNN fails to combine the variation of NN search and does not focus on spatial data with low dimensionality.

The algorithm derived from narrative pruning techniques answer the probabilistic RNN queries on multidimensional uncertain data. However, it fails to provide solution the probabilistic Rk-NN queries on uncertain data. Although RNN focused on discrete case, pruning rules are used when the uncertain objects are denoted using probability density function [1]. Probabilistic Inverted (PI) index uses the probability density function efficiently. PI index calculates the lower bound and upper bound for a threshold keyword query through which incompetent nodes are pruned and experienced nodes are returned as early as possible. PI index finds all quasi-SLCA results meeting the threshold need. Though, it fails to extend the technique on tackling the correlation [4].

General optimal route query derived from backward and forward search are used to react with a variation of the optimal route queries. Routes require covering subset of the specified categories though scalability issue arises on the optimal route query [6]. Decision Tree Classification in [5] is primarily designed to handle uncertain data with remarkably higher accuracies. Decision tree building algorithms controls data tuples with uncertain values in higher cost than through processing single values. Averaging approach and Distribution-based approach is a combination of mathematical theorems. The theorems also allocates essential pruning of large search space with split point determination throughout tree construction. Decision tree construction on uncertain data contains the inconsistency in CPU utilization.

III. SPACE QUERY CLASSIFIER INDEXING FOR MINING UNCERTAIN DATA

Classification is a classical issue in machine learning and data mining. Data uncertainty is an inherent property in various applications because of outdated sources or imprecise measurement. When data mining techniques are used in the data, their uncertainty obtains high quality results. In a set of training data tuples, each tuples contains a class label and it is denoted by a feature vector. The main objective is to design a model that predicts the class label of an unseen test tuple based on the tuple's feature vector. A method for managing the data uncertainty is to reduce the probability distributions by their statistics like means and variances. Indexing, multidimensional uncertain objects, range query and processing are the results in many applications such as information cleaning, Radio Frequency Identification (RFID) networks, Location-Based Services (LBS), Global Position System (GPS), Sensor Data Analysis, Economic Decision making and market surveillance.

A. Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data

Reverse Nearest Neighbor (RNN) query is an important query type in many areas. Uncertain data are inbuilt in many applications like sensor databases, moving object databases, market analysis, and quantitative economic research. In these applications, the exact values of data are indefinite because of restriction of evaluating equipment, delayed data updates, incompleteness, or data anonymization to protect privacy. In general, uncertain object is denoted in two methods: by probability density function (i.e. continuous case) and by all probable cases with an allocated probability (i.e., discrete case).

In figure 1, probabilistic RNN queries are described with an example. Consider the probabilistic RNN Query has three residential blocks A, B, and Q. The houses within each block are denoted by small circles. The centroid of each residential block is denoted as a hollow circle. For privacy reasons, the residential blocks where the people live however fails to have information about accurate addresses of their houses. The probability is allocated to all possible location of a person with the residential block. The exact location of a person in A is a1 with 0.5 possibilities. Conventional queries on residential blocks use the distance functions by calculating the distance between the centre points of two blocks. The issues of probabilistic RNN queries on uncertain data are

explained by the semantics of potential worlds. A new probabilistic RNN query processing framework uses many novel pruning approaches with the probability threshold and geometric, topological and metric properties. In addition, highly optimized verification method is based on upper and lower bounding of the RNN probability of candidate objects. Many applications for the queries take the proximity of uncertain objects and the applications of RNNs on uncertain objects are similar. Probabilistic RNN query processing has new demands in planning the new efficient algorithms. RNN query processing derived from pruning methods is studied. The pruning methods fail to relate with probabilistic RNN queries or changed into ineffective. Uncertain objects contain arbitrary shapes of the uncertain regions. In addition, pruning rules are connected with the occurrence level of uncertain objects in expensive manner because each uncertain object contains large number of occurrences.

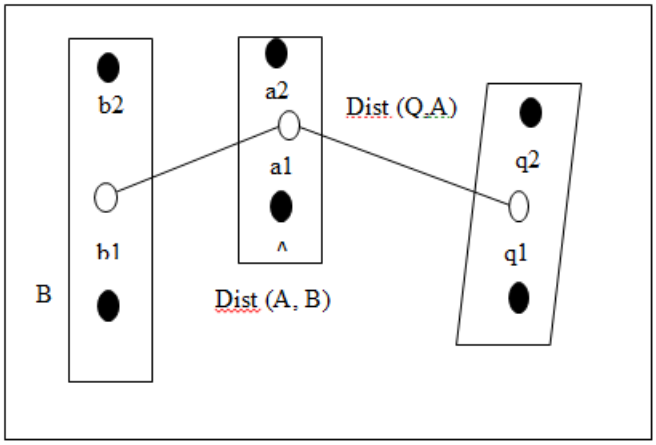

Figure 1: Probabilistic RNN Query

B. Quasi-SLCA based Keyword Query Processing over Probabilistic XML Data

Uncertainty is used in many web applications like information extraction, information integration and web data mining. In uncertain database, probabilistic threshold queries are studied where all results satisfy the queries with possibilities equal to or larger than the threshold values.

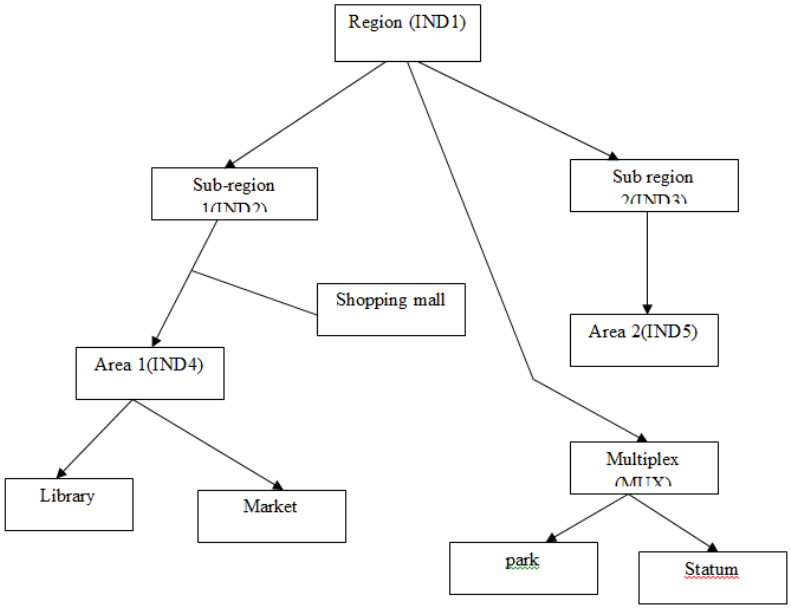

Figure 2: Probabilistic XML Data Tree

As the flexibility of XML data model allocates a natural demonstration of uncertain data, uncertain XML data management contains significant problem. A new ProbabilisticThreshold Keyword Query (PrTKQ) over uncertain XML databases derived from quasi-SLCA semantics is planned. In general, an XML document is

observed as a rooted tree. In the rooted tree, each node denotes an element or contents. XIRQL maintains keyword search in XML derived from structured queries. The users fail to contain the knowledge of the structure of XML data or the query language. The LCA-based approaches initially recognize the LCA node which has the every keyword under the sub tree at least once. A PrXML document describes a probability distribution over a space of deterministic XML documents. Each deterministic document with the space is called a possible world

A PrXML document represented as a labeled tree has ordinary and distributional nodes. Ordinary nodes are regular XML nodes and they emerge in deterministic documents. The distributional nodes are used for probabilistic process of creating deterministic documents in efficient manner and fail to happen in PrXML documents. A new probabilistic XML model called PrXML {IND,MUX} is designed. A PrXML document is taken as a labeled tree in which distributional nodes includes two types, IND and MUX. An IND node has children that are independent of each other. The children of a MUX node are mutually-exclusive. One child are present in a random instance document is termed as a possible world. A real number from [0, 1] is joined on each edge in the XML tree with the conditional probability where the child node emerges under the parent node specified the presence of the parent node.

C. Decision Trees for Uncertain Data

Classification is a classical problem in machine learning and data mining. One of the classification models is the decision tree model. Decision trees are well developed one as they are practical and easy to understand. Rules are removed from decision trees. Many algorithms like ID3 and C4.5 are used for decision tree construction. The algorithms are applied in many applications like image identification, medical analysis, target marketing, scientific tests, fraud detection, and credit rating of loan applicants. In traditional decision tree classification, a feature of a tuple is definite or mathematical. An exact and distinct point value is taken. In many applications, data uncertainty is frequent one. The value of a feature/attribute is captured not using a single point value. It is captured using range of values to a probability distribution.

A simple method is used to control data uncertainty is to abstract the probability distributions through summary statistics like means and variances called Averaging. An additional approach is to obtain information by the probability distributions for designing a decision tree called Distribution-based. The issues of designing the decision tree classifiers on data with uncertain numerical attributes are solved. The key aim is to plan an algorithm for constructing the decision trees from uncertain data by Distribution-based approach. It also studies whether the Distribution-based approach results in higher classification accuracy than the Averaging approach. It is employed to create a theoretical foundation where the pruning techniques are derived. In addition, it increases the computational efficiency of the Distribution-based algorithms.

A simple method is designed to manage the uncertain information and also replaces each pdf with the expected value. It also changes the data tuples into point-valued tuples. It also minimizes the issues back for point valued data. The traditional decision tree algorithms like ID3 and C4.5 are reprocessed called Averaging (AVG). It utilizes an algorithm derived fromC4.5.

IV. PERFORMANCE ANALYSIS OF SPACE QUERY CLASSIFIER INDEXING FOR MINING UNCERTAIN DATA

In order to compare the space query classifier indexing for mining uncertain data using different techniques, number of queries is taken to perform the experiment. Various parameters are used for query classification of uncertain data.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Query processing efficiency is defined based on the queries addressed to the total number of queries by the user with different interval of time periods. It is measured in terms of percentage (%).

$$QueryProcessingEfficiency\ (\%) = \frac{Queriesaddressed}{Totalnumberofqueries} * Time * 100 \quad (1)$$

Query Processing Efficiency comparison takes place on existing Probabilistic XML Model (PrXML), Decision Tree Classification and Probabilistic Reverse Nearest Neighbor (RNN) Query Processing Framework.

Table 1. Tabulation of query processing efficiencyfor query classifier indexing for mining uncertain data

Number of Queries (Number)	**Query Processing Efficiency (%)**		
	PrXML Model	***Decision Tree Classification***	***Probabilistic RNN Query Processing Framework***
10	65	56	61
20	69	59	64
30	72	62	68
40	76	66	73
50	81	69	77
60	86	74	81
70	89	78	85

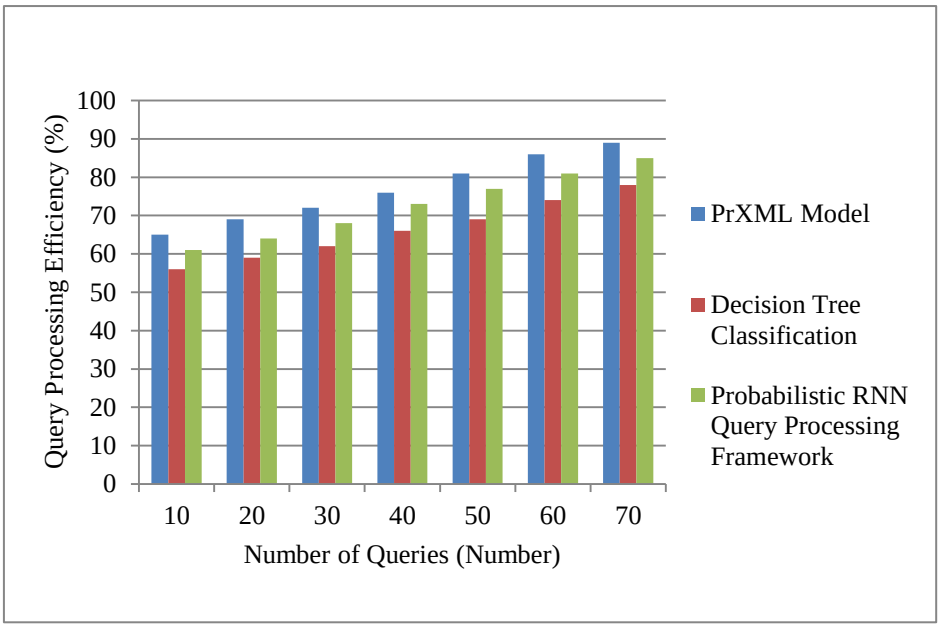

Figure 3: Query Processing Efficiency for Query Classifier Indexing for Mining Uncertain Data

From figure 3, query processing efficiency for query classifier indexing for mining uncertain data are calculated. Probabilistic XML Model (PrXML) is higher query processing efficiency than that of Decision Tree Classification and Probabilistic Reverse Nearest Neighbor (RNN) Query Processing Framework. Research in Probabilistic XML Model (PrXML) has13.78% higher query processing efficiency than, Decision Tree Classification and 5.45% higher query processing efficiency than Probabilistic Reverse Nearest Neighbor (RNN) Query Processing Framework.

B. Execution Time

Execution time is defined as the difference between starting time and ending time of query classification of uncertain data. It is measured in terms of millisecond (ms).

$$ExecutionTime\ (ms) = StartingTime - Endingtimeofqueryclassification \quad (2)$$

Execution time comparison takes place on existing Probabilistic XML Model (PrXML), Decision Tree Classification and Probabilistic Reverse Nearest Neighbor (RNN) Query Processing Framework.

From figure 4, execution time for query classifier indexing for mining uncertain data are calculated. Decision Tree Classification consumes less amount time than that of Probabilistic XML Model (PrXML) and Probabilistic Reverse Nearest Neighbor (RNN) Query Processing Framework. Research in Decision Tree

Classification consumes 36.05% less amount of time than Probabilistic XML Model (PrXML) and 63.81% lesser execution time than Probabilistic Reverse Nearest Neighbor (RNN) Query Processing Framework. Tabulation of execution time for query classifier indexing for mining uncertain data.

Table 2. Tabulation of execution time for query classifier indexing for mining uncertain data

Number of Queries (Number)	Execution Time(ms)		
	PrXML Model	***Decision Tree Classification***	***Probabilistic RNN Query Processing Framework***
10	45	24	36
20	49	27	39
30	53	31	43
40	57	35	48
50	59	39	52
60	64	43	55
70	69	48	58

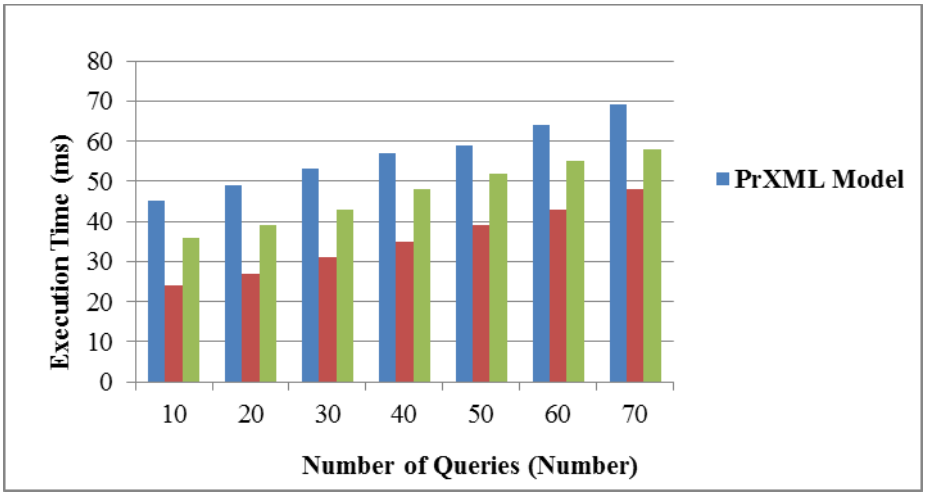

Figure 4: Execution Time for Query Classifier Indexing for Mining Uncertain Data

C. Memory Consumption

Memory consumption is defined as the amount of memory space consumed while classifying the queries on uncertain data. It is measured in terms of MegaBytes (MB).

$$MemoryConsumption(MB) = Totalmemoryspace - unusedmemoryspace \quad (3)$$

Table 3. Tabulation of memory consumption for query classifier indexing for mining uncertain data

Number of Queries (Number)	Memory Consumption (MB)		
	PrXML Model	***Decision Tree Classification***	***Probabilistic RNN Query Processing Framework***
10	18	23	13
20	21	28	16
30	25	34	19
40	29	39	22
50	34	43	25
60	38	46	28
70	42	50	32

Memory Consumption comparison takes place on existing Probabilistic XML Model (PrXML), Decision Tree Classification and Probabilistic Reverse Nearest Neighbor (RNN) Query Processing Framework.

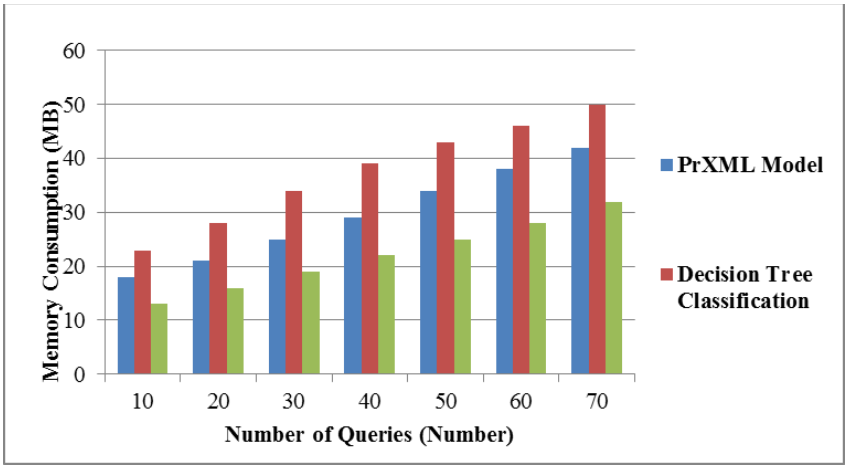

Figure 5: Memory Consumption for Query Classifier Indexing for Mining Uncertain Data

From figure 5, memory consumption for query classifier indexing for mining uncertain data are evaluated. Probabilistic Reverse Nearest Neighbor (RNN) Query Processing Framework consumes less memory space than that of Probabilistic XML Model (PrXML) and Decision Tree Classification. Research in Probabilistic Reverse Nearest Neighbor (RNN) Query Processing Frameworkconsumes33.72% lesser memory space than Probabilistic XML Model (PrXML) and 71.52% lesser memory than Decision Tree Classification.

V. Discussion on Limitation of Space Query Classifier Indexing for Mining Uncertain Data

Probabilistic reverse nearest neighbor query recovers the objects from uncertain data with high probability. The threshold is provided to RNN of uncertain query object. Nontrivial pruning rules are designed for uncertain data and for the probability threshold. Probabilistic RNN queries are the verification of candidate objects that incurs substantial cost. However, the solution to probabilistic RkNN queries on uncertain data is not provided. Probabilistic inverted (PI) index quickly return the qualified answers and filter out the unqualified ones based on lower/upper bounds. Quasi-SLCA results in threshold keyword over uncertain XML data. It also satisfies the possible semantics employed from the assumption of Gaussian distribution. Probability value for each keyword in a node are calculated and stored in PI index. It also fails to extend the technique to tackle the correlations. The correlations take place on certain data points.

Decision tree classification is designed to handle uncertain data. It is useful for designing decision trees using classical algorithms when there are tremendous amount of data tuples. Inconsistency-tolerant repair are occurred. It also fails in problematic use of triggers and the non-declarative design is reduced in favor of ITIC.

A. Related Works

The existing data inference and compression substrate over RFID streams as illustrated in [8] takes a time-varying graph model to collect the probable object locations and inter object connections. RFID streams with probabilistic algorithm calculate the location and containment for each object but not inefficient form of query processing in distributed environments. Pareto-Based Dominant Graph (DG) is planned in offline mode to clear the dominant relationship between records. Custer-based storage scheme reduces I/O cost in Traveler algorithm. However, it fails to relate the DG index in dominant relationship analysis [9].

Dynamic programming algorithm in PTIME Categorize lineages of IQ queries into path lineages and composite lineages to minimize the path length. PTIME Greedy algorithm computes approximate responsibilities but fails to compute responsibility for unions of conjunctive queries [7].

B. Future Direction

The future direction of query classifier indexing for mining uncertain data can be carried out through nearest neighbor approximation techniques to decrease the execution time on classification. In addition, space query indexing is used on uncertain data to reduce the overhead count. Further classification of uncertain data using conjunctive query generator over a simple subclass of conjunctive queries solves the tolerance problem.

VI. CONCLUSION

In this article, space query classifier indexing for mining uncertain data is studied. Classification is a classical issue in machine learning and data mining. Data uncertainty is an inherent property in various applications because of outdated sources or imprecise measurement. Reverse Nearest Neighbor (RNN) query is an important query type in many areas. From the survey, solution to probabilistic RkNN queries on uncertain data is not provided. The wide range of experiments on existing techniques calculates the comparative results of the various query classification techniques and its limitations. Finally from the limitation identified from the existing works, further research work can be carried out with space query indexing and query generator for reducing the tolerance problem and also reduces the overhead count.

REFERENCES

[1] Muhammad Aamir Cheema., Xuemin Lin., Wei Wang., Member., Wenjie Zhang., and Jian Pei., "Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data," IEEE transactions on knowledge and data engineering, Vol. 22, No. 4, April 2010

[2] Sze Man Yuen., Yufei Tao., Xiaokui Xiao., Jian Pei., and Donghui Zhang., "Superseding Nearest Neighbor Search on Uncertain Spatial Databases," ieee transactions on knowledge and data engineering, Vol. 22, No. 7, July 2010

[3] Tobias Farrell., Kurt Rothermel., and Reynold Cheng., "Processing Continuous Range Queries with Spatiotemporal Tolerance," ieee transactions on mobile computing, Vol. 10, No. 3, March 2011

[4] Jianxin Li., Chengfei Liu., Rui Zhou., and Jeffrey Xu Yu., "Quasi-SLCA based Keyword Query Processing over Probabilistic XML Data," January 2013

[5] Smith Tsang., Ben Kao., Kevin Y. Yip., Wai-Shing Ho., and Sau Dan Lee., "Decision Trees for Uncertain Data," ieee transactions on knowledge and data engineering, Vol. 23, No. 1, January 2011

[6] Jing Li., Yin David Yang., and Nikos Mamoulis., "Optimal Route Queries with Arbitrary Order Constraints," ieee transactions on knowledge and data engineering, Vol. 25, No. 5, May 2013

[7] Biao Qin., Shan Wang., Xiaofang Zhou., Xiaoyong Du., "Responsibility Analysis for Lineages of Conjunctive Queries with Inequalities," IEEE transactions on knowledge and data engineering., 2013

[8] Yanming Nie., Richard Cocci., Zhao Cao., Yanlei Diao., and Prashant Shenoy., "SPIRE: Efficient Data Inference and Compression over RFID Streams," IEEE transactions on knowledge and data engineering, Vol. 24, No. 1, January 2012

[9] Lei Zou., and Lei Chen., "Pareto-Based Dominant Graph: An Efficient Indexing Structure to Answer Top-K Queries,"IEEE transactions on knowledge and data engineering, 2011.

[10] Hendrik Decker., and Davide Martinenghi., "Inconsistency-Tolerant Integrity Checking,"IEEE transactions on knowledge and data engineering, Vol. 23, No. 2, February 2011.

BIOGRAPHY

M. Kalavathi has received the M.Sc., Degree from Bharathiar University in 2004, M.Phil Degree from Annamalai University in 2006, respectively in Computer Science. Her research interest includes Data Mining and Digital Image Processing.

Dr. P. Suresh has received the M.Sc., Degree from Bharathidasan University in 1995, M.Phil Degree from Manonmaniam Sundaranar University in 2003, M.S (By Research) Degree from Anna University, Chennai in 2008, PGDHE Diploma in Higher Education and Ph.D., Degree from Vinayaka Missions University in 2010 and 2011 respectively in Computer Science. He is an Editorial Advisory Board Member of Elixir Journal. His research interest includes Data Mining and Natural Language Processing. He is a member of Computer Science Teachers Association, New York.