Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 9, September 2013, pg.205 – 211

RESEARCH ARTICLE

Well-organized Estimation of Range Aggregates against Uncertain Location-Based Queries

G.Nithya[1], M.Chinnusamy[2]

[1]Research Scholar, Department of Computer Science, Vivekanandha College, Elayampalayam, Tiruchengode, Tamil Nadu, India

[2]Assistant Professor, Department Of Computer Science, Vivekanandha College, Elayampalayam, Tiruchengode, Tamil Nadu, India

[1] nithya.it22@gmail.com; [2] chithrachinnu@rediffmail.com

ABSTRACT: - We consider the problem of efficient estimation of distance between uncertain objects. In many real-life applications, data such as sensor readings and weather forecasts are usually uncertain when they are collected or produced. An uncertain object has a probability distribution function (PDF) to represent the probability that it is actually located in a particular location. A fast and accurate distance computation between uncertain objects is important to many uncertain query evaluation (e.g., range queries and nearest-neighbor queries) and uncertain data mining tasks (e.g., classifications, clustering and outlier detection). However, existing approaches involve distance computations between samples of two objects, which is very computationally intensive. On one hand, it is expensive to calculate and store the actual distribution of the possible distance values between two uncertain objects. On the other hand, the expected distance (the weighted average of the pair wise distances among samples of two uncertain objects) provides very limited information and also restricts the definitions and usefulness of queries and mining tasks.

Keywords- Uncertainty; index; range aggregate; query Attacks

I. INTRODUCTION

Sensor networks and image processing produce uncertain data that, with the recent growth in these activities, is of interest to researchers working on how to support various kinds of interesting queries and data mining of these data. While there has been a large amount of research work done on mining and queries on relational databases, the focus has been on databases that store data in exact values. In many real-life applications, however, the raw data such as sensor data are usually uncertain when they are collected or produced. Sources of uncertain data include readings from sensors, information extracted using probabilistic parsing of input sources, classification results of image processing using statistical classifiers, results from predictive programs used for the stock market, and weather forecasts in meteorology. These uncertain data may be in the form of an exact value with margins of error, sometimes with or without a probability Distribution (or density) function. The result may also be represented as an interval or a set of values, one of which may be the real value. However, since traditional databases store only exact values, uncertain data are usually transformed into exact data by, for example, taking the weighted average or mean value (for quantitative attributes) or by taking the value with the highest frequency or possibility. This makes the storage, query and mining much simpler because it allows the use of existing commercial database systems and mining techniques, but there is an obvious shortcoming: By approximating the uncertain source data values, the intermediate and final results from the mining tasks and queries will also be approximate and may be wrong. For example, the locations of cancroids of clusters deviate from the real ones; errors may appear in the calculation of distances between objects; or some data may be even assigned to the wrong clusters.

The distance between two data objects is a measurement used in various queries and data mining tasks such as nearest-neighbor queries and clustering (e.g., K-means clustering [1]). While it is very simple to calculate the distance between two exact data objects by applying a distance formula, it is not trivial when the two data objects' locations are uncertain. An uncertain object has more than one possible location. If each object oi has ni possible locations, then we have n1n2 possible distances between objects o1 and o2 for every possible pairwise combination of their locations. Given a probability distribution Pi of the possible locations of object oi, we can calculate a probability distribution of the possible distances. This result is very informative but it is very expensive to compute, especially when the number of possible locations is infinite. Instead, as in [2], an expected distance is used which calculates the average of all the possible distances between samples weighted by their probabilities. Recent research related to data mining on uncertain data such as [2], and [3] obtains the expected distance by assuming that the information of the precise probabilities of all possible locations are known in advance. The information is either represented as (i) a discrete probability distribution function (PDF) where probabilities are given on the finite set of possible locations, or (ii) a continuous probability distribution function (or probability density function, where the probability density is defined on a region.

II.RELATED WORK

Query imprecision or uncertainty may be often caused by the nature of many applications, including location based services. The existing techniques for processing location-based spatial queries regarding certain query points and data points are not applicable or inefficient when uncertain queries are involved. In this paper, we investigate the problem of efficiently computing distance based range aggregates over certain data points and uncertain query points as described in the abstract. In general, an uncertain query Q is a multidimensional point that might appear at any location x following a probabilistic density function pdfðxÞ within a region Q: region. There are a number of applications where a query point may be uncertain. Below are two sample applications. A blast warhead carried by a missile may destroy things by blast pressure waves in its lethal area where the lethal area is typically a circular area centered at the point of explosion (blast point) with radius [10] and depends on the explosive used. While firing such a missile, even the most advanced laser-guided missile cannot exactly hit the aiming point with 100 percent guarantee. The actual falling point (blast point) of a missile blast warhead regarding a target point usually follows some probability density functions (PDFs); different PDFs have been studied in [24] where vicariate normal distribution is the simplest and the most common one [10]. In military applications, firing such a missile may not only destroy military targets but may also damage civilian objects .

Therefore, it is important to avoid the civilian casualties by estimating the likelihood of damaging civilian objects once the aiming point of a blast missile is determined. If q1 in Fig. 1 is the actual falling point of the missile, then objects p1 and p5 will be destroyed. Similarly, objects p2, p3, and p6 will be destroyed if the actual falling point is q2. In this application, the risk of civilian casualties may be measured by the total number n of civilian objects which are within distance away from a possible blast point with at least probability. Note that the probabilistic threshold is set by the commander based on the levels of tradeoff that she wants to make between the risk of civilian damages and the effectiveness of military attacks; for instance, it is unlikely to cause civilian casualties if n ¼ 0 with a small .Moreover, different weight values may be assigned to these target points and hence the aggregate can be conducted based on the sum of the values.

A. Problem Definition

If we view an attribute as a dimension, then the union of the domains of all attributes produces a multidimensional space where a certain object is represented as a point. Due to the uncertain nature or actual system limitation in the data collection phase, the imperfect data quality leads to uncertain attribute values of an object. Therefore an uncertain object may be represented as a set of points, each of which is a possible location of the object. A discrete probability distribution function is used to represent the distribution of the probabilities of the possible locations. Alternatively, an uncertain object may also be represented as a (finite or infinite) region, which covers the possible locations of the object (especially when the number of possible locations is not finite). We call this region the uncertainty domain of object oi, denoted as UD(oi).A continuous probability distribution function (or probability density function, pdf), pi, is used to indicate the probability density of each possible location x within the region.

III. COMBINED DNA, RNA COMBINED WITH PCR TECHNIQUE

A. Sampling and DNA extraction

To achieve the best results in a PCR assay it is crucial to take particular care during sampling. Conventional culture-dependent microbial techniques require aseptic condition during sampling and transportation and less time possible is needed before starting analyses. The advantage of PCR approach is that analyses can be performed also in a different day from sampling, using the precaution of freezing samples at -20 °C in order to avoid any loss of DNA quality and any growth of micro organisms in the sample during storage.

This procedure is not appropriate when a pre-enrichment cultivation step is needed before PCR, in order to augment the number of target cells. In our experience the best way to facilitate sampling and to enhance DNA amplification is by storing the sample in sterile dark plastic bags and freezing it immediately while sampling, if possible. Otherwise, the use or refrigerated bags (4 °C) is opportune; in that case it's recommended to start with DNA extraction within few hours from sampling. When sampling water, it could be very useful to pre-filter the sample on nitrocellulose sterile filters (0,45 μm or 0,2 μm) and then freezing and storing the filter as a starting matrix for DNA extraction.

B. PCR Technique

The aim of this study was to adapt this novel technique for use in the rapid laboratory-based detection of *Plasmodium* spp. and to validate the sensitivity of this technique in comparison to that of conventional nested PCR and microscopy Patient samples were collected between 2007 and 2009 at the MARIB (Malaria Research Initiative Bandarban) center in Bandarban, Chittagong Hill Tracts, Bangladesh, as part of a hospital- and field-based fever survey. Written informed consent was obtained from all study participants or their legal representatives, and the study protocol were approved by the appropriate ethical review committee.

From all participating patients aged 8 years and older, 100 μl venous blood was drawn. From patients younger than 8 years, 2 drops of blood obtained by finger prick was collected and transferred onto 903 filter

paper (Schleicher & Schuell BioScience GmBH, Dassel, Germany) in duplicate. Filter papers were air dried at room temperature and stored under airtight conditions at 4°C until further processing. A total number of 140 filter paper samples are included in the evaluation.

C. RNA Method

RNA replication is the copying of one RNA to another. Many viruses replicate this way. The enzymes that copy RNA to new RNA, called RNA-dependent RNA polymerases, are also found in many eukaryotes where they are involved in RNA silencing.[4] RNA editing, in which an RNA sequence is altered by a complex of proteins and a "guide RNA", could also be considered an RNA-to-RNA transfer.

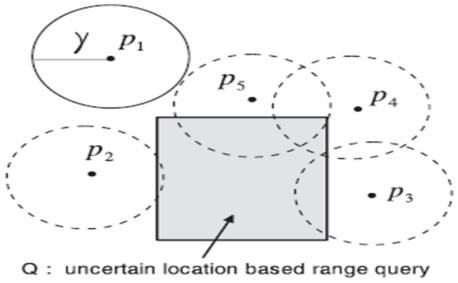


Fig.1 Example uncertain location based range query

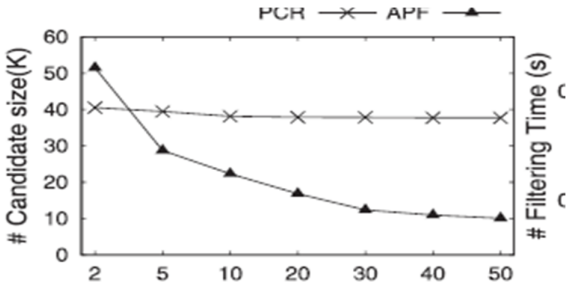
IV. EXPERIMENT RESULTS

We present results of a comprehensive performance study to evaluate the efficiency and scalability of the techniques proposed in the paper. Following the frame work of Algorithm 1 in Section 3, four different filtering techniques have been implemented and evaluated. MMD. The maximal/minimal distance filtering technique is employed as a benchmark to evaluate the efficiency of other filtering techniques. The statistical filtering technique proposed in Section 3.2.. PCR. The PCR technique discussed in Section 3.3. For the fairness of the comparison, we always choose a number of PCRs for the uncertain query such that space.

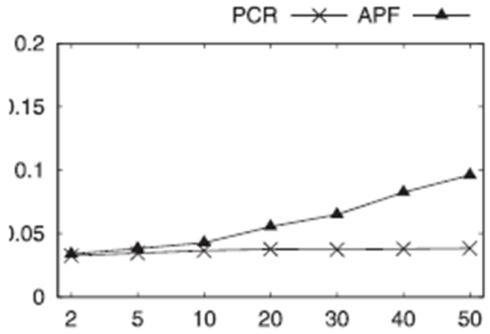
TABLE 1

SYSTEM PARAMETER

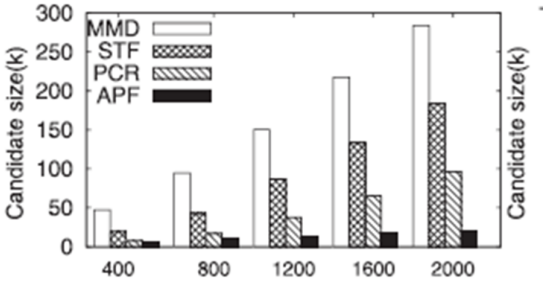
Notation	**Definition (Default Values)**
qr	The radious of the uncertain region (600)
σ	Standard deviation for Normal distribution (300)
N_{ap}	The number of anchor points in APF (30)
N_d	The size of Da for each anchor point (30)
γ	Query distance (1200)
θ	Probabilistic threshold (∑[0,1])
N	The number of data points in p (1m)



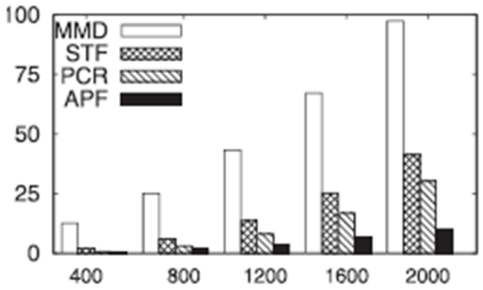
Fig,2 Diff. *Nap*



Fig,3 Diff. *Nap*



Fig,4 US



Fig, 5 3D Uniform Points

The IO cost is the number of pages visited from RP, and the candidate size is the number of points which need verification. In addition, we also evaluate the filtering time and the query response time. In each

experiment, we issue 200 uncertain queries and use the average candidate size, number of nodes accessed, filtering time, and query response time to evaluate the performance of the techniques.

Lists parameters which may potentially have an impact on our performance study. In our experiments, all parameters use default values unless otherwise specified.

A. Performance Evaluation

In this section, we conduct comprehensive experiments to evaluate the effectiveness and efficiency of our filtering techniques proposed in the paper. In the first set of experiments, we evaluate the impact of query distance on the performance of the filtering techniques in terms of candidate size, IO cost, query response time, and filtering cost. All evaluations are conducted against US and 3d Uniform data sets. Fig. 14 reports the candidate size of MMD, STF, PCR, and APF when query distance _ grows from 400 to 2,000. We set dap to 30 and 60 for US and 3d Uniform data sets, respectively. As expected, larger value results in more candidate data points in the verification phase. It is interesting to note that with only a small amount of statistic information, DNA can significantly reduce the candidate size compared with MMD. PCR can further reduce the candidate size while RNA significantly outperforms others especially for the large only 19,650 data points need to be further verified for APF when 2; 000 on US data set, which is 96,769, 1, 83,387 and 284,136 for PCR, RNA, and DNA, respectively.

V. CONCLUSION

We have described the limitation of expected distance, and the importance of calculating and representing the distance distribution between an uncertain object pair in queries and data mining applications. We have proposed two Gaussian approximation approaches to represent the actual distance distribution, which can be calculated and stored efficiently and effectively. it practical for the research communities to define and develop more powerful queries and data mining tasks based on the distance distribution (rather than just the expected distance). With only two parameters (mean and variance), the storage space of the distance pdf can be reduced significantly.

REFERENCES

[1] P.K. Agarwal, S.-W. Cheng, Y. Tao, and K. Yi, "*Indexing Uncertain Data*," Proc. Symp. Principles of Database Systems (PODS), 2009.
[2] C. Aggarwal and P. Yu, "*On High Dimensional Indexing of Uncertain Data*," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), 2008.
[3] C. Bohm, M. Gruber, P. Kunath, A. Pryakhin, and M. Schubert, "*Prover: Probabilistic Video Retrieval using the Gauss-Tree*," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), 2007.
[4] C. Bohm, A. Pryakhin, and M. Schubert, "*Probabilistic Ranking Queries on Gaussians*," Proc. 18th Int'l Conf. Scientific and Statistical Database Management (SSDBM), 2006.
[5] J. Chen and R. Cheng, "*Efficient Evaluation of Imprecise Location-Dependent Queries*," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE),2007.
[6] R. Cheng, J. Chen, M.F. Mokbel, and C.-Y. Chow, "*ProbabilisticVerifiers: Evaluating Constrained Nearest-neighbor Queries over Uncertain Data*," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2008.
[7] R. Cheng, D.V. Kalashnikov, and S. Prabhakar, "*Evaluating Probabilistic Queries over Imprecise Data*," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2003.
[8] R. Cheng, S. Singh, and S. Prabhakar, "*Efficient Join Processing over Uncertain Data*," Proc. Int'l Conf. Information and Knowledge Management (CIKM), 2006.
[9] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J.S. Vitter, "*Effcient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data*," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2004.
[10] G.M. Siouris, "*Missile Guidance and Control Systems*". Springer Publication, 2004.
[11] X. Dai, M. Yiu, N. Mamoulis, Y. Tao, and M. Vaitis, "*Probabilistic Spatial Queries on Existentially Uncertain Data*," Proc. Int'l Symp. Large Spatio-Temporal Databases (SSTD), 2005.
[12] E. Frentzos, K. Gratsias, and Y. Theodoridis, "*On the Effect of Location Uncertainty in Spatial Querying*," IEEE Trans. Knowledge Data Eng., vol. 21, no. 3, pp. 366-383, Mar. 2009.
[13] M. Hua, J. Pei, W. Zhang, and X. Lin, "*Ranking Queries on Uncertain Data: A Probabilistic Threshold Approach*," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2008.
[14] Y. Ishikawa, Y. Iijima, and J.X. Yu, "Spatial *Range Querying for Gaussian-Based Imprecise Query Objects*," Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), 2009.
[15] H.-P. Kriegel, P. Kunath, M. Pfeifle, and M. Renz, "*Probabilistic Similarity Join on Uncertain Data*," Proc. Int'l Conf. Database Systems for Advanced Applications (DASFAA), 2006.

[16] H.P. Kriegel and M. Pfeifle, "*Density-Based Clustering of Uncertain Data*," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), 2005.
[17] X. Lian and L. Chen, "*Monochromatic and Bichromatic Reverse Skyline Search over Uncertain Databases*," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2008.
[18] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "*Efficient Clustering of Uncertain Data*," Proc. Int'l Conf. Data Mining (ICDM), 2006.
[19] M.A. Soliman, I.F. Ilyas, and K.C. Chang, "*Top-k Query Processing in Uncertain Databases*," Proc. Int'l Conf. Data Eng. (ICDE), 2007.
[20] Y. Tao, R. Cheng, X. Xiao, W.K. Ngai, B. Kao, and S. Prabhakar, "*Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions*," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2005.

