Research Journal of Applied Sciences 9 (8): 511-517, 2014
ISSN: 1815-932X

Uncertain Data Retrieval Using Contour Cluster Vector Technique in Single Dimensional Database

[1]M. Sathish and [2]S. Sukumaran
[1]Department of Computer Application Sree Amman Arts Science,
[2]Department of Computer Science, Erode Arts and Science College, Erode, Tamil Nadu, India

Abstract: For the past decade, the usage of d-dimensional databases in applications is considerably increasing in several fields like medicine, biology and CAD/CAM applications. The most popular research now running on is the retrieval of similar set of uncertain objects from the single-dimensional databases through indexing. However, for retrieving the relative set of uncertain objects from the database, searching plays a basic functionality. Several researchers implemented diverse set of approaches for similarity search so-called feature transformation. As a result, the similarity search is changed into a look for points in the feature space which are close to a particular query point in the multi-dimensional feature space. Numerous index structures and algorithms have been planned. It has been revealed that the process of presenting the novel index structures significantly develop the presentation in indexing the single-dimensional databases. In this research, researchers first made a focus on faster retrieval of data from single dimensional database. Contour-Cluster Vector (CCV) technique is formed by creating cluster that contains time series subsequences of approximately the same contour (shape). For each cluster, a lower-bound distance is computed for the users' query and the most similar element of the cluster. CCV demonstration can be used as index for single dimensional data and that it permits more efficient similarity search. Simulation experiments conducted with a large number of instances per object to evaluate the efficacy of proposed CCV technique against indexing the multi-dimensional uncertain objects through range searching. The performance of the proposed CCV technique is evaluated in terms of processing time, number of instantiations and throughput and the evaluation results showed that it achieved 10-12% high in attaining the retreival of data.

Key words: Single-dimensional data, uncertain objects, contour cluster vector technique, indexing, similarity search

INTRODUCTION

With a growing number of novel database applications for instance multimedia content-based application, time series and systematic databases, the plan of indexing and query processing methods over high dimensional data sets turn out to be a significant research area. These requests utilize the so-called feature revolution which converts significant features or properties of data objects into high-dimensional points, i.e., each characteristic vector comprises of set of dimensional values which communicate to coordinates in a given dimensional space. Penetrating the objects with these features is, therefore, a look for points in this feature space. In given d-dimensional databases, indexes are crucial to uphold the queries in multi-dimensional database:

- K-Nearest Neighbor (KNN) queries: discover the K-most related objects in the database relating to a particular object
- Similarity, range queries: discover all objects in the database which are inside a specified distance from a point
- Window queries: discover all objects whose attributes lies inside a specified ranges

There is an extensive stream of study on resolving the similarity search setback and numerous multidimensional indexes have been planned. But these index formations have basically been considered in the perspective of disk-based systems where it is understood that the databases are too huge to form into the central memory. This supposition is gradually more being confront as RAM gets cheaper and better. This includes

Corresponding Author: M. Sathish, Department of Computer Application Sree Amman Arts Science, Erode, Tamil Nadu, India

encouraged interest in study in main memory databases. Conventionally, Clustering algorithms manage a set of objects whose locations are precisely identified and do not deal with condition in which object positions are indecisive. Data uncertainty is nevertheless, inbuilt in numerous real-life applications owing to factors for instance the arbitrary character of the corporeal data creation and compilation processes, dimension errors and data staling.

Indexing d-dimensional data has been the center of attention of a substantial study of research over several years but no normally decided pattern has appeared to process with the blow of the B-Tree for instance, on the indexing of one-dimensional data. Simultaneously, the requirement of retrieving the same set of data in an environment where databases turn out to be larger and more intricate in their construction for extracting the important information become more complicated.

Zhang *et al.* (2012) presented effectively indexing multi-dimensional uncertain objects for range searching which investigate extension of indexing technique and plan to tackle correlation among multidimensional uncertain objects.

Now-a-days, a number of oblique data compilation methodologies have guide to the propagation of tentative data. Such, data points are habitually symbolized in the structure of a probabilistic function as the resultant deterministic value is not recognized. This enhances the confronts over mining and running uncertain data, since the specific activities of the basic data is no longer recognized (Wang *et al.*, 2011). This research gap, motivated us to develop an CCV technique to tackle the faster retrieval of data among multi-dimensional uncertain objects.

Literature review: While the determined data of numerous database applications such as OLAP and scientific studies are distinguished by very high dimensionality, characteristic queries posed on these data demand to a diminutive number of applicable dimensions. Unfortunately, the multi-dimensional access techniques considered for high-dimensional data achieve rather defectively for the specified queries. Ngai *et al.* (2011) considered the crisis of clustering data objects with position uncertainty. In a representation, a data object is symbolized by an improbability section over which a probability density function (pdf) is clear. One technique to group such indecisive objects is to concern the UK-Means algorithm, an expansion of the conventional K-Means algorithm which consigns every object to the group whose representative has the minimum distance from it.

The objective of top-k ranking for objects is to grade the objects so that the finest k of them can be dogged. Aggarwal and Yu (2009) measured an object to be a unit which comprises a number of attributes whose job in the object is dogged by an aggregation utility. For indecisive data, the semantic base of top-k objects turns out to be imprecise. The semantics is designed and modeled by indecisive data where the standards of an object's characteristics are articulated by probability distributions and controlled by some declared conditions.

The study (Tuncel *et al.*, 2004) examined the association among rate-distortion assumption and proficient content-based data reclamation from high-dimensional databases. The design of the database is considered as the programming a data object series and reclamation from the database as the deciphering of the series using side information (i.e., the query) accessible only at the decoder. Even though numerous data structures have been presented for position and spatial indexing, none of them is recognized to key points in both overlapping and non-overlapping sections in a finest query recovery time. The study (AnandhaKumar *et al.*, 2010) introduced the cross tree-a multidimensional data formation for indexing two dimensional position spaces.

Given a set of multi-dimensional points, the skyline holds the finest points along with any utility that is monotone on all axes. Tao *et al.* (2006) presented a method SUBSKY which resolves the crisis utilizing a single B-tree and can be realized in any relational database. The central part of SUBSKY is a revolution that changes multi-dimensional data to 1D value and allows numerous efficient pruning heuristics.

Now-a-days, the particular spatial index formation can not assure the outsized scale three-dimensional spatial data association and administration. Considering these features, a novel spatial index formation named LOD_SKDR tree construction is considered for running of large scale spatial database which is united with SKD-Tree, R-Trees and LOD object information (Liu *et al.*, 2010).

Constructing a search engine to recognize a related content in the World-wide web needs: robust video relationship dimensions; rapid similarity search methods on large databases and instinctive association of search results. Cheung and Zakhor (2005) analyzed the remaining two subjects by presenting a feature extraction system for rapid similarity search and a clustering algorithm for recognition of parallel clusters.

Lehman *et al.* (2008) presented a similarity-based Searching and Pattern Matching algorithm that recognizes time series data with parallel sequential dynamics in large-scale databases. A time series sections is

symbolized by feature vectors that reproduce the dynamical prototypes of single and multi-dimensional physiological time series. Existing approaches selling with this crisis rely on a particular similarity measure among graph composition. Abbaci *et al*. (2011) recommended an unusual approach for searching related graphs to a graph query where comparison among graphs is rather represented by a vector of scalars than a exclusive scalar.

In explorative data examination, the data underneath deliberation habitually exists in in a High-Dimensional (HD) data space. At present numerous methods are accessible to examine this kind of data. So far, automatic approaches comprise dimensionality diminution and cluster analysis, whereby visual-interactive techniques aim to offer efficient visual mappings to demonstrate, transmit and plot a route HD data (Tatu *et al*., 2012).

High-dimensional indexing techniques have been established relatively valuable for response time expansion. Based on euclidian distance, several papers have been discussed with the applications where data vectors are high-dimensional. But, these techniques do not usually maintain similarity search for data retrieval. To enhance the data retrieval process, in this research, CCV technique is implemented with the d-dimensional database.

MATERIALS AND METHODS

To enhance the process of retrieving the uncertain data objects from the single dimensional database, in this research, contour vector cluster technique is presented. The CCV technique analyzes the similar elements in the database and clustered based on the time series subsequences. After formation of cluster, the lower bound values are determined based on the users' query and the evaluation of similar element of the clustering scheme is done. Finally, similarity search is performed to determine the required set of data objects from the database. The process of the proposed CCV technique is presented in Fig. 1.

In a single dimensional database, uncertain set of objects are placed at diverse set of locations in the specified region. Each uncertain data objects are associated with the probability density function and the region where the uncertain object actually resides. To retrieve the similar set of uncertain data objects from the database, clustering is done based on the subsequences of time series. After formation of cluster, similarity search is done at the clustered set to access the items in an efficient manner.

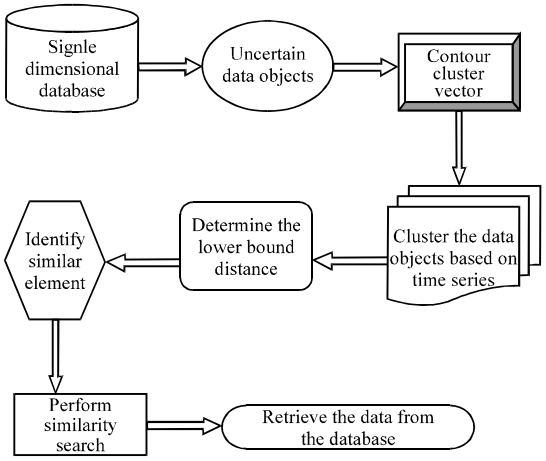

Fig. 1: Architecture diagram of the proposed CCV technique

Contour cluster vector technique: For a given dimensional database, the uncertain set of data objects are analyzed and processed first. At first, the specific uncertain data object values are analyzed. Before analyzing, a limit value L has been set for clustering processes. If the data object lies within a limit, form a group with a limited range values. Or else, form a group with restricted set of attribute values. Consider a set of data objects which are represented in a sequential format as $S = \{d_1, d_2, ..., d_n\}$. Identify the length of each sequence of data objects in the given database. Based on the time series length, the sequences are evaluated. The distnace between the two diverse sequences are determined based on euclidean distance which is represented as:

$$ED(S1,S2) = \sqrt{\sum_{i=1}^{L} (S1_i - S2_i)^2} \quad (1)$$

Where:
S1 and S2 = Set of sequences
L = Data object limit value

Similarity, search is done based on the process that identifies set of sequences S2 whose Euclidean distances to the sequence S1 are lies within a specified distance. Usually, the data objects in the d-dimensional database are represented in two diverse formats:

- Whole matching
- Subsequence matching

Whole matching contains a set of sequence as of data objects with a same length and the subsequence matching has different value in length. Here, based on the

time series subsequences, i.e., data is represented as a series of numbers, the clustering is done among the data objects. Identify the attribute value of each uncertain data objects. Before identification, set some limit to the range value. According to that, clustering is done with the time series subsequences of data objects. The algorithm describes the process of contour vector clustering technique for clustering the time series subsequences database. The formation of cluster is shown in Fig. 2.

// Algorithm:

Input: Time series subsequences data objects, attribute value, limit value L, max time T, processing time t
Begin
Set L = α
Analyze the uncertain data objects in the given database D
Form a sequences of data objects based on real numbers
Identify the longer sequence of data object
For each uncertain data object
    Identify its attribute value (V)
    Do
    If (V lies with α)
Identify the position of the data object
Form a group
Else
Search its V after a while
End If
While (t met with T)
End

Based on the above procedures followed, the clustering takes place with the uncertain data objects. To made the faster retrieval of information, researchers perform indexing scheme to enhance the process. Figure 2 describes the process of describing the cluster formation for the specified time series. It revealed that the same cluster part varies based on the time series. At a particular time, the value lies within a range and in some time it exceeds the limit. At this stage, the process of retrieval of data from the dataset is complex to proceed. To resolve this issue in cluster formation for retrieval of data, indexing structure is used which is described in the study.

Index construction: After formation of cluster, the part of each cluster is analyzed to identify the set of similar elements in the database. A pattern of bits is represented for every cluster formation in the database. It decides the process of choosing the exact precise bits for the clsuter in forms. After all set of series of data objects are fomred inside a cluster, comparison is made inside a cluster itself to determine the similar element level over the cluster set C. This process is referred as indexing where the unique subsequences of the data objects are labeled. Each cluster has one or more subsequences of the uncertain data objects which has similar set of attirbute values which also includes the distance among every pair of subsequence in that cluster. The construction of indexing based on data objects is defined as:

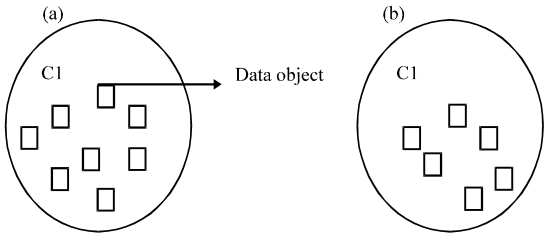

Fig. 2: a) Cluster part at a time series t1 and b) Cluster part at a time series t2

$$I(D)=\sum_{i=0}^{n}\frac{N}{Cap(C)}\prod_{j=0}^{n}(\sigma_j+\varepsilon) \quad (2)$$

Where:

I = Indexing structure
D = Data objects
N = Number of sequences of data
C = Cluster part
Cap(C) = Capacity of cluster
ε = Distance measured for a given query sequence

$$\sigma_j=\frac{Cap(C)}{N} \quad (3)$$

The process of constructing the index for the specified set of data objects in dimensional database is described in the algorithm.

// Algorithm:

Input: Set of time series sequences S, Elements n, data object D = {d_1, d_2, ..., d_n} distance measured ε
Analyze the set of sequences in the database stored in a clustered format
Access each set of data sequence from the database
Sort the given set of data objects in the clustered part based on its attribute value
For each d_n
Build the index based on its attribute value
Based on users' query,
Indexing access is performed using Eq. 2
Process all the query sequences Q
End For
End

For a given set of users' query, the retrieval of information and sorting is done based on the diverse kinds and it is specified as:

- Entire cluster searching
- Cluster object ordering
- Intra cluster searching

In entire cluster searching process, the processes of identifying the similar set of elements are determined based on the users' query. If the cluster does not have similar set of subsequences, the particular cluster is

eliminated from the list. Entire-clustering searching process searches the entire clusters for the data based on users' query and make use of the triangular variation to reduce the repeated set of items in a cluster once researchers have observed the presence of more similar items inside the cluster.

In cluster object ordering, the ordering of the clustered data objects are done based on best first order. Usually, the set of clusters taken for searching at first is processed with the best matching sequences. After formation of indexing clustered part, the lower bound between the query and the sequences are denoted as:

$$\min\_dis(R,BS) = \sum_i |R_i| \times ag(BS_i R_i)$$
$$ag(BS_i R_i) = \begin{cases} 1 & \text{build the index} \\ 0 & \text{else} \end{cases} \quad (4)$$

Where:
R = Query sequence
BS = Time series subsequence value
Ag() = Determine the similar element in the cluster

This expression permits us to promptly compute the best cluster order in which to provide the set of required uncertain data objects to the search algorithm.

In intra cluster searching, the clustered data objects are processed at the intra cluster level. Based on users' query, similarity search is done at the intra cluster level and the set of necessary items are retrieved from the dataset. The intra cluster searching for the retrieval of data is done similar to the tin tire cluster searching process but it followed the lower bound distance scheme to select the items from the cluster. The performance of the proposed CCV technique is presented as:

- Clustering efficiency
- Processing time
- Information retrieval rate

Clustering efficiency measures the effectiveness of clusters based on the attribute values for processing the users' queries of diverse lengths.

Processing time measures the time taken to process the user queries of different lengths and it are referred in terms of seconds. Information retrieval rate defines the rate at which the required information is retrieved based on the users' requests for the given dataset.

Experimental evaluation: An experimental evaluation is carried out to estimate the performance of the proposed CCV technique with the dataset extracted from UCI repository. The diabetes dataset is taken for the evaluation of experiments to analyze the results. The diabetes dataset is described in Table 1. Diabetes patient records were attained from two sources:

Table 1: Dataset description

Dataset characteristics	Multivariate, time-series
Attribute characteristics	Categorical, Integer
Number of attributes	20
Number of web hits	68241

- Automatic electronic recording device
- Study records

The automatic device contained a domestic clock to timestamp events while the study records only afforded reasonable time slots (breakfast, lunch, dinner, bedtime). For study records, predetermined times were dispersed to breakfast (08:00), lunch (12:00), dinner (18:00) and bedtime (22:00). Therefore, study records contain consistent recording times while electronic records contain more sensible time stamps. Diabetes files consist of four fields per record. Each field is divided by a label and each record is alienated by a novel line. The attributes are described as:

Date in MM-DD-YYYY format:

- Time in XX: YY format
- Code
- Value

The code field is deciphered as follows:
33 = Regular insulin dose
34 = NPH insulin dose
35 = Ultra Lente insulin dose
48 = Unspecified blood glucose measurement
57 = Unspecified blood glucose measurement
58 = Pre-breakfast blood glucose measurement
59 = Post-breakfast blood glucose measurement
60 = Pre-lunch blood glucose measurement
61 = Post-lunch blood glucose measurement and so on

RESULTS AND DISCUSSION

In this study, the performance of the proposed CCV technique and the existing range search in multi dimensional database is evaluated. Figure 3 describes the evaluation results of the proposed CCV technique for data retrieval.

The clustering efficiency is measured based on the size of available data in the dataset. The value of the proposed CCV technique is compared with the existing range searching is illustrated in Fig. 3.

Figure 3 describes the clustering efficiency which is measured based on the size of available data in the dataset. Compared to the existing range searching technique, the proposed CCV technique provides high

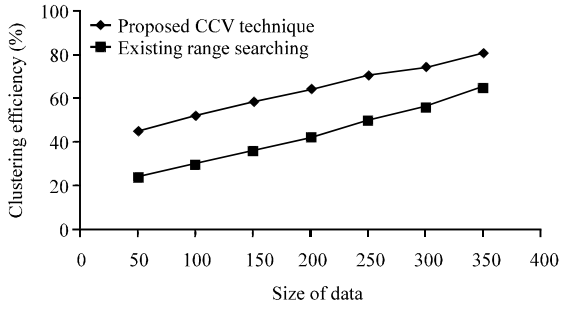

Fig. 3: Size of data vs. clustering efficiency

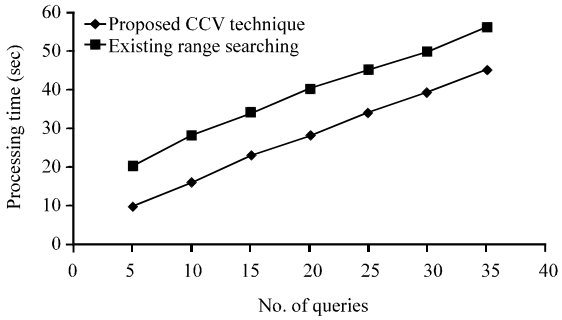

Fig. 4: Number of queries vs. processing time

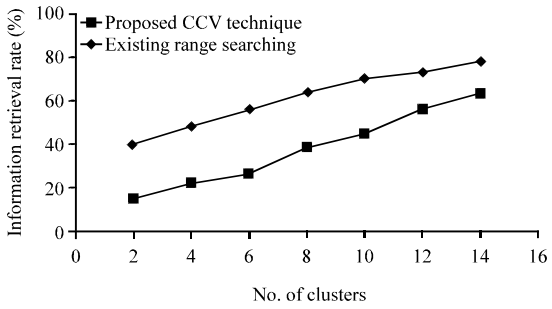

Fig. 5: Number of clusters vs. information retrieval rate

efficiency in clustering the data. Because, the proposed CCV clustered the data based on its attribute value. A range has been set for clustering the uncertain data objects. The clustering capacity changes based on the time series subsequences of data in the dataset. But, in the Existing Range Searching Method, a Cost Model is represented for indexing technique which does not provide effective clustering part. The variance in the clustering efficiency is 15-20% high in the proposed CCV technique.

The processing time is measured based on the number of queries sent by the user for the given dataset. The value of the proposed CCV technique is compared with the existing range searching is illustrated in Fig. 4.

Figure 4 describes the processing time which is measured based on the users; request queries to the database. Compared to the existing range searching method, the proposed CCV technique consumes less processing time for processing the queries. Since, the proposed CCV technique followed the clustering process, the processing time of incoming queries are done in a short interval of time. So, the proposed CCV technique supports processing the more number of queries in the database. The variance in the processing time is 5-11% high in the proposed CCV technique.

The information retrieval rate is measured based on the number of clusters in the given dataset. The value of the proposed CCV technique is compared with the existing range searching is illustrated in Fig. 5.

Figure 5 describes the information retrieval rate which is measured based on the number of clusters in the given dataset. Compared to the existing range searching method, the proposed CCV technique achieves high information retrieval rate. Since, the proposed CCV technique followed indexing framework, the similarity search over the set of data is done effectively. Proposed technique supports three kinds of searching techniques, the retrieval of data is fast. The variance in the retrieval rate is 10-15% high in the proposed CCV technique.

Finally, it is being observed that the proposed CCV technique provides high information retrieval rate by implementing the contour vector cluster technique. The CCV technique provides a framework for building the indexing scheme by determining the lower bound distance among the set of uncertain data objects.

CONCLUSION

In this research, a novel technique is presented to enhance the retrieval of information from the database by implementing the contour cluster vector technique. Based on the attribute values of the uncertain data objects, the clustering takes place with the dataset. The clustering is done based on varied set of time series and the subsequences of the dataset are clustered according to that. This increases the complexity of similarity searching in the large set of database. To resolve this issue, the proposed CCV technique constructed the indexing framework. To make the data search in the clustered dataset more efficient, indexing scheme is used with three diverse techniques: entire cluster searching, intra level clustering and the cluster object ordering.

At first, the indexing is done based on the indexing access cost for the uncertain data objects and then the similarity search is proceeded. After that the techniques

are implemented by determining the lower bound distance among the subsequences of dataset. Finally, the search is accomplished by assigning the CCV technqiue as index. The main advantage of contour vector clustering is followed similarity search outcome is accurate since the lower bound distance has been determined with the given dataset.

Experimental evaluation is conducted with the diabetes dataset extracted from UCI repository for the estimating the performance of the proposed CCV technique. Performance results revealed that the proposed CCV technique provides higher rate in information retrieval and efficiency, low processing time.

REFERENCES

Abbaci, K., A. Hadjali, L. Lietard and D. Rocacher, 2011. A similarity skyline approach for handling graph queries: A preliminary report. Proceedings of the 27th International Conference on Data Engineering Workshops, April 11-16, 2011, Hannover, Germany, pp: 112-117.

Aggarwal, C.C. and P.S. Yu, 2009. A survey of uncertain data algorithms and applications. IEEE Trans. Knowl. Data Eng., 21: 609-623.

AnandhaKumar, P., J. Priyadarshini, C., Monisha K. Sugirtha and S. Raghavan, 2010. Location based hybrid indexing structure-R kd tree. Proceedings of the 1st International Conference on Integrated Intelligent Computing, August 5-7, 2010, Bangalore, India, pp: 140-145.

Cheung, S.C. and A. Zakhor, 2005. Fast similarity search and clustering of video sequences on the world-wide-web. IEEE Trans. Multimedia, 7: 524-537.

Lehman, L.H., M. Saeed, G.B. Moody and R.G. Mark, 2008. Similarity-based searching in multi-parameter time series databases. Proceedings of the Conference on Computers in Cardiology, September 14-17, 2008, Bologna, Italy, pp: 653-656.

Liu, Y., G. Liu and Z. He, 2010. Spatial index technology for multi-scale and large scale spatial data. Proceedings of the 18th International Conference on Geoinformatics, June 18-20, 2010, Beijing, China, pp: 1-4.

Ngai, W.K., B. Kao, R. Cheng, M. Chau, S.D. Lee, D.W. Cheung and K.Y. Yip, 2011. Metric and trigonometric pruning for clustering of uncertain data in 2D geometric space. Inform. Syst., 36: 476-497.

Tao, Y., X. Xiao and J. Pei, 2006. Subsky: Efficient computation of skylines in subspaces. Proceedings of the 22nd International Conference on Data Engineering, April 3-7, 2006, Atlanta, GA., USA., pp: 65-65.

Tatu, A., F. Maas, I. Farber, E. Bertini, T. Schreck, T. Seidl and D. Keim, 2012. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. Proceedings of the IEEE Conference on Visual Analytics Science and Technology, October 14-19, 2012, Seattle, WA., USA., pp: 63-72.

Tuncel, E., P. Koulgi and K. Rose, 2004. Rate-distortion approach to databases: Storage and content-based retrieval. IEEE Trans. Inform. Theory, 50: 953-967.

Wang, C., L.Y. Yuan and J.H. You, 2011. On the semantics of top-k ranking for objects with uncertain data. Comput. Mathe. Appl., 62: 2812-2823.

Zhang, Y., W. Zhang, Q. Lin and X. Lin, 2012. Effectively indexing the multi-dimensional uncertain objects for range searching. Proceedings of the 15th International Conference on Extending Database Technology, March 27-30, 2012, Berlin, Germany, pp: 504-515.