On High Dimensional Indexing of Uncertain Data

Charu C. Aggarwal, Philip S. Yu

IBM T. J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532, USA
{ charu, psyu }@us.ibm.com

***Abstract*— In this paper, we will examine the problem of distance function computation and indexing uncertain data in high dimensionality for nearest neighbor and range queries. Because of the inherent noise in uncertain data, traditional distance function measures such as the L_q-metric and their probabilistic variants are not qualitatively effective. This problem is further magnified by the sparsity issue in high dimensionality. In this paper, we examine methods of computing distance functions for high dimensional data which are qualitatively effective and friendly to the use of indexes. In this paper, we show how to construct an effective index structure in order to handle uncertain similarity and range queries in high dimensionality. Typical range queries in high dimensional space use only a subset of the ranges in order to resolve the queries. Furthermore, it is often desirable to run similarity queries with only a subset of the large number of dimensions. Such queries are difficult to resolve with traditional index structures which use the entire set of dimensions. We propose query-processing techniques which use effective search methods on the index in order to compute the final results. We discuss the experimental results on a number of real and synthetic data sets in terms of effectiveness and efficiency. We show that the proposed distance measures are not only more effective than traditional L_q-norms, but can also be computed more efficiently over our proposed index structure.**

I. Introduction and Overview

In recent years, many advanced technologies have been developed to store and record large quantities of data continuously. In many cases, the data may contain errors or may be only partially complete. For example, sensor networks typically create large amounts of uncertain data sets. In other cases, the data points may correspond to objects which are only vaguely specified, and are therefore considered uncertain in their representation. Similarly, surveys and imputation techniques create data which is uncertain in nature. This has created a need for *uncertain data management* algorithms and applications.

In uncertain data management, data records are represented by probability distributions rather than deterministic values. Therefore, a data record is represented by the corresponding parameters of a multi-dimensional probability distribution. Some examples in which uncertain data management techniques are relevant are as follows:

- The uncertainty may be a result of the limitations of the underlying equipment. For example, the output of sensor networks is often uncertain. This is because of the noise in sensor inputs or errors in wireless transmission.
- In many cases such as demographic data sets, only partially aggregated data sets are available. Thus, each aggregated record is actually a probability distribution.
- In privacy-preserving data mining applications, the data is perturbed in order to preserve the sensitivity of attribute values. In some cases, probability density functions of the records may be available.
- In some cases, data attributes are constructed using statistical methods such as forecasting or imputation. In such cases, the underlying uncertainty in the derived data can be estimated accurately from the underlying methodology.

The problems of distance function computation and indexing are closely related, since the construction of the index can be sensitive to the distance function. Furthermore, effective distance function computation is inherently more difficult in the high dimensional or uncertain case. Direct extensions of distance functions such as the L_q-metric are not very well suited to the case of high dimensional or uncertain data management. This is because these distances are most affected by the dimensions which are most dissimilar. In the high dimensional case, the statistical behavior of the sum of these dissimilar dimensions leads to the *sparsity problem*. This results in similar distances between every pair of points, and the distance functions are often *qualitatively ineffective* [12]. Furthermore, the dimensions which contribute most to the distance between a pair of records are also likely to have the greatest uncertainty. Therefore, the effects of high dimensionality are magnified by the uncertainty, and the contrast in distance function computations is lost. The challenge is to design a distance function which continues to be both *qualitatively effective* and *index-friendly*.

The problem of indexing has been studied extensively in the literature both for the case of deterministic data [5], [6], and for the case of uncertain data [8], [9], [17], [18]. However these techniques do not deal with some of the unique challenges in the similarity indexing of high dimensional or uncertain data. These unique challenges are as follows:

- Similarity functions need to be carefully designed for the high dimensional and uncertain case in order to maintain contrast in similarity calculations. Furthermore, the distance function needs to be sensitive to the use of an index.
- In most cases, the similarity or range queries are only performed on a small subset of the dimensions of high dimensional data. For example, in many applications,

978-1-4244-1837-4/08/$25.00 © 2008 IEEE

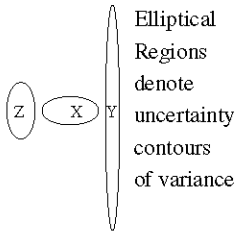

Fig. 1. Effects of uncertainty on distance

we are likely to perform a range query only over 3 to 4 dimensions of a 100-dimensional data set. Such queries cannot be processed with the use of traditional index structures, which are designed for full-dimensional queries.

The queries which can be resolved with the use of our index structure are as follows:

- Determine the nearest neighbor to a given target record in conjunction with an effective distance function.
- Determine the nearest neighbor to the target T by counting the expected number of dimensions for which the points lie within *user-specified* threshold distances $t_1 \ldots t_d$. In practical applications, only a small number of the threshold values $t_1 \ldots t_d$ may be specified, and the remaining are assumed to be ∞. In such cases, the nearest neighbor search is performed only over a small number of projected dimensions which are relevant to that application.
- For a given subset of dimensions S, and a set of ranges $R(S)$ defined on the set S, determine the points which lie in $R(S)$ with probability greater than δ. We note that this particular query is referred to as a *projected range query*, since we are using only a subset of the dimensions.

We will see that the key is to construct a distance function which can be computed efficiently in the high dimensional case, and is both qualitatively effective and index-friendly. We will refer to this index structure as UniGrid (or UNcertain Inverted GRID Structure).

Our approach for indexing involves two steps:

- The construction of a distance function which works effectively in high dimensionality, and continues to be efficient.
- The design of an index which works effectively with this distance function.

The construction of the index uses a thresholding method which can design the distance functions more effectively in the high dimensional case, by computing the expected number of dimensions for which the distance is less than a certain threshold.

This distance function is constructed in conjunction with an index which uses a two-level inverted representation. The two-level inverted representation can access the data both by locality and variance. This index is used in conjunction with novel query processing approaches in order to perform the indexing. More details on the approach may be found in [1].

REFERENCES

[1] C. C. Aggarwal, P. S. Yu: On High Dimensional Indexing of Uncertain Data. *IBM Research Report*, 2007.

[2] C. C. Aggarwal, P. S. Yu: The IGrid Index: Reversing the Dimensionality Curse in High Dimensional Space. *ACM KDD Conference*, 2000.

[3] C. C. Aggarwal. On Density Based Transformations for Uncertain Data Mining. *ICDE Conference*, 2007.

[4] D. Barbara, H. Garcia-Molina, D. Porter: The management of probabilistic data. *IEEE Transactions on Knowledge and Data Engineering*, 4(5), pp. 487–502, 1992.

[5] N. Beckmann, H.-P. Kriegel, R. Schneider, B. Seeger. The R^*-Tree: An Efficient and Robust Access Method for Points and Rectangles. *ACM SIGMOD Conference*, 1994.

[6] S. Berchtold, D. Keim, H.-P. Kriegel. The X-Tree: An Index Structure for High Dimensional Data. *VLDB Conference*, 1996.

[7] D. Burdick, P. Deshpande, T. S. Jayram, R. Ramakrishnan, S. Vaithyanathan. OLAP Over Uncertain and Imprecise Data, *VLDB Conference* pp. 970–981, 2005.

[8] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, J. Vitter: Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data. *VLDB Conference*, 2004.

[9] R. Cheng, D. Kalashnikov, S. Prabhaker: Evaluating Probabilistic Queries over Imprecise Data. *SIGMOD Conference*, 2003.

[10] N. Dalvi, D. Suciu: Efficient Query Evaluation on Probabilistic Databases. *VLDB Conference*, 2004.

[11] A. Das Sarma, O. Benjelloun, A. Halevy, J. Widom: Working Models for Uncertain Data. *ICDE Conference*, 2006.

[12] A. Hinneburg, C. Aggarwal, D. Keim. What is the nearest neighbor in high dimensional space. *VLDB Conference*, 2000.

[13] H.-P. Kriegel, M. Pfeifle: Density-based clustering of uncertain data. *KDD Conference*, 2005.

[14] L. V. S Lakshmanan, N. Leone, R. Ross, V. S. Subrahmanian. ProbView: A Flexible Database System. *ACM TODS*, 22(3):419–469, 1997.

[15] S. I. McClean, B. W. Scotney, M. Shapcott. Aggregation of Imprecise and Uncertain Information. *IEEE TKDE*, 13(6):902–912, 2001.

[16] D. Pfozer, C. Jensen. Capturing the uncertainty of moving object representations. *SSDM Conference*, 1999.

[17] S. Singh, C. Mayfield, S. Prabhakar, R. Shah, S. Hambrusch: Indexing Uncertain Categorical Data. *ICDE Conference*, 2007.

[18] Y. Tao, R. Cheng, X. Xiao, W. Ngai, B. Kao, S. Prabhakar. Indexing Multi-dimensional Uncertain Data with Arbitrary Probability Density Functions. *VLDB Conference*, 2005.