

A Resourceful Algorithm for High Dimensional Spaces

¹Aretty Narayana, ²M.Srinivasa Rao, ³Dr. R.V.Krishnaiah

^{1,2}Dept. of CSE, DRK College of Engineering & Technology, Hyderabad, AP, India

³Dept of CSE, DRK Group of Institutions, Hyderabad, Andhra Pradesh, India

Abstract

The approaches for resemblance search in correlated, high dimensional data sets are derived with in a clustering framework. To conflict the “Curse of Dimensionality” a technique namely Vector approximation by indexing was proposed, employs scalar quantization and necessarily ignores dependencies across dimensions and a source of sub optimality is represented. In contrast inter dimensional correlations are exploited by clustering and thus a more packed in representation of data set. However, existing methods reduce the irrelevant clusters which are based on bounding hyper spheres and bounding rectangles which lack their tightness and compromises the efficiency in nearest neighbor search. In this paper, based on separating hyper plane boundaries of Voronoi clusters we propose a new cluster adaptive distance bound to complement our cluster based index. With a relatively small preprocessing storage overhead a bound is enabled by an efficient spatial filtering and is applicable to similarity measures. Experiments show that our indexing method is scalable with data set size and dimensionality and outperforms with several recent proposed indexes.

Keywords

High Dimensional, Clusters, Hyper Plane, Indexing

I. Introduction

There has been an increase of personal digital media devices with the development in semiconductor technology and powerful signal processing tools [1-2], such as digital cameras, music and digital video players. Similarly, the storage media have become cheaper to manufacture and correspondingly their capacities have increased. This has created new applications that can store large amounts of data periodically in and later, and can retrieve it from databases. The size of these databases can range from the relatively small to the very large database. In the future, for various purposes such as data mining and decision support, large organizations will have to retrieve and process peta bytes of data. And thus, there exist numerous applications that can access large multimedia databases. In high-dimensional spaces, the spatial queries, especially nearest neighbor queries have been studied widely and several analyses have concluded that the nearest neighbor search, due to the notorious “curse of dimensionality” it is impractical at high dimensions with Euclidean distance metric, and others have suggested that this may be over pessimistic [3-5]. Specifically, the search performance is the intrinsic dimensionality of the data set and not the dimensionality of the address space. In many previous studies the typical assumption is that with independent attributes, the data is uniformly distributed and such data-sets have been shown to exhibit the “curse of dimensionality” in that, distance between all pairs of points i.e. in high dimensional spaces converges to the same value is shown in fig. 1 .

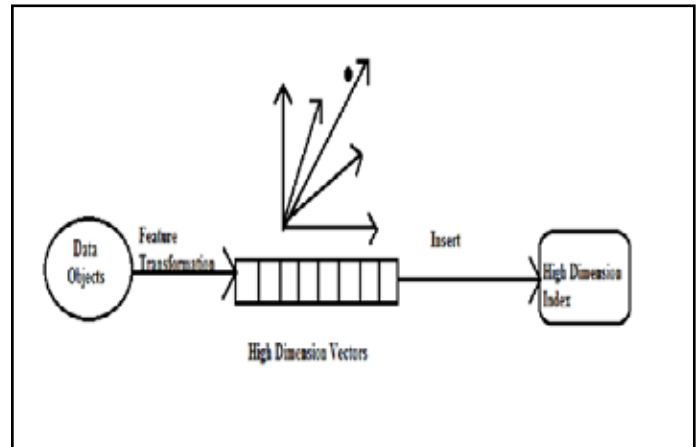


Fig. 1: High Dimension Indexing Diagram

Clearly, to index the nearest and furthest neighbors such data types are impossible and are indistinguishable [6-7]. However, they are typically skewed and exhibit intrinsic dimensionalities that are much lower than their embedding dimension as the real data sets overwhelmingly invalidate assumptions of independence and/or uniform distributions; rather due to subtle dependencies between attributes and real data-sets are demonstrably indexable with Euclidean distances and it is entirely different matter if euclidean distance is perceptually acceptable [8-9]. Hence, we mainly focus on real data-sets and compare performance against the state-of-the-art indexing targeting real data-sets.

II. Correlated Work on Adaptive Clustering

The index structures survive and facilitate search and retrieval of multi-dimensional data and the recursive partitioning of the space with a combination of hyper-spheres and hyper-rectangles in low dimensional spaces or hyper-rectangles, hyper-spheres, have been found to be effective for nearest neighbor search and retrieval [10-11]. For metric spaces with arbitrary distance functions the Euclidean distance and the Metrics (M-trees) have been found to be effective with the previous methods. Such multi-dimensional indexes work well with low dimensional spaces and they outperform sequential scan. The performance degrades with increase of feature dimensions and, after a certain dimension threshold; it becomes inferior to sequential scan which has been observed in this paper. Such performance degradation is attributed to Bellman's 'curse of dimensionality' which refers to the exponential growth of hyper-volume with dimensionality of the space [12].

III. Estimation of Vector Files

To overcome the curse of dimensionality, a popular and effective technique i.e. the vector approximation files (VA-File) is used. To obtain a quantized estimate for the data that reside inside the cells VA-File partitions the space into hyper-rectangular cells [13]. The vector approximation file is sequentially scanned and upper and lower bounds on the distance from the query vector to each cell are estimated during nearest neighbor search, as a non-empty cell locations are encoded into bit strings and are stored in a separate approximation file on the hard disk. The final set of candidate vectors are then read from the hard disk and the

exact nearest neighbors are determined and the bounds are used to prune irrelevant cells. Each component of the feature vector is separately and uniformly quantized as the terminology "Vector Approximation" is actually performed as scalar quantization. IDistance and Pyramid Tree are other techniques which are based on local dimensionality reducing transformations [14-15]. Typically the centroid, are evaluated as the data-set is partitioned and, the distances of the resident vectors to some reference point. By using the ubiquitous one-dimensional indexes such as the B+-tree the feature vectors in a partition are now indexed by their centroid-distance. Spheres of gradually increasing radii are drawn around the query during query processing, until they intersect a cluster sphere and the relevant elements in the partition are identified by centroid-distances which lie in the intersecting region, are retrieved for finer scrutiny. The feature vectors and distance functions are often only approximations of user perception of similarity. Hence, with additional rounds of query refinement are necessary even the results of an exact similarity search are inevitably perceptually approximate. The reader is directed for a more detailed survey of approximate similarity search. The optimal tradeoffs between search quality and search time has also been studied within an information theoretic framework.

IV. Results

Various extensive experiments were conducted on real data-sets to differentiate and evaluate the performance of index structure with other popular methods. A variety of solutions, in terms of random vs. sequential input and outputs was obtained for each index by varying the number of clusters (or some appropriate parameters), and the performance of the curves were consequently compared.

V. Conclusion

Significant correlations and non-uniform distributions are exhibited by real multidimensional data-sets and indexing is suboptimal by performing uniform, scalar quantization. In this paper, based upon principles of vector quantization instead we have proposed an indexing method, where the data set is partitioned into Voronoi clusters and the clusters are accessed in the order of the query-cluster distances. Based on separating hyper plane boundaries and our search index, which are complemented by these bounds, and are developed by a cluster distance bounds, which is applicable to Euclidean distance metrics. When allowed the same number of sequential pages, it has obtained significant reductions in number of random inputs and outputs over several recently proposed indexes, and has a low computational cost and scales well with dimensions and size of the data-set. The hyper plane bounds are better than minimum bounding hyper rectangle (MBR) and minimum bounding hyper sphere (MBS) bounds, and they are still loose compared with the true query-cluster distance. Possibly by optimizing the clustering algorithm so as to optimize the cluster distance bounds to be further tightened.

References

- [1] C. Faloutsos, "Searching in Multimedia Databases", By Content. Kluwer Academic Press, 1996.
- [2] K. S. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, "When is "nearest neighbor" meaningful?", in ICDT, 1999, pp. 217–235.
- [3] C. C. Aggarwal, A. Hinneburg, D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces", in ICDT, 2001, pp. 420–434.
- [4] B. U. Pagel, F. Korn, C. Faloutsos, "Deflating the dimensionality curse using multiple fractal dimensions", in ICDE, 2000, pp. 589–598.
- [5] T. Huang, X. S. Zhou, "Image retrieval with relevance feedback: From heuristic weight adjustment to optimal learning methods", in ICIP, Vol. 3, 2001, pp. 2–5.
- [6] J. Davis, B. Kulis, P. Jain, S. Sra, I. Dhillon, "Information-theoretic metric learning", in ICML, 2007, pp. 209–216.
- [7] R. Weber, H. Schek, S. Blott, "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces", in VLDB, August 1998, pp. 194–205.
- [8] A. Gersho, R. M. Gray, "Vector Quantization and Signal Compression. Kluwer Academic Publishers, 1992.
- [9] T. Zhang, R. Ramakrishnan, M. Livny, "BIRCH: An efficient data clustering method for very large databases", in SIGMOD, 1996, pp. 103–114.
- [10] N. Katayama, S. Satoh, "The SR-tree: An index structure for highdimensional nearest neighbor queries", in SIGMOD, May 1997, pp. 369–380.
- [11] D. A. White, R. Jain, "Similarity indexing with the SS-tree", in ICDE, 1996, pp. 516–523.
- [12] N. Beckmann, H. Kriegel, R. Schneider, B. Seeger, "The R*-tree: an efficient and robust access method for points and rectangles", in SIGMOD, 1990, pp. 322–331.
- [13] Y. Sakurai, M. Yoshikawa, S. Uemura, H. Kojima, "The A-tree: An index structure for high-dimensional spaces using relative approximation", in VLDB, September 2000, pp. 516–526.
- [14] S. Berchtold, C. Bohm, H. V. Jagadish, H. P. Kriegel, J. Sander, "Independent Quantization: An index compression technique for highdimensional data spaces," in ICDE, 2000, pp. 577–588.
- [15] C. Yu, B. C. Ooi, K. L. Tan, H. V. Jagadish, "Indexing the distance: An efficient method to knn processing", in VLDB, September 2001, pp. 421–430.



Aretty Narayana has completed B.Tech (E.C.E) from Srinivas Reddy Institute of Technology and pursuing M.Tech (C.S.E) in DRK College of Engineering and Technology, JNTUH, Hyderabad, Andhra Pradesh, India. His main research interest includes Data Mining, Database and DWH.



M. Srinivasa Rao is working as an Associate Professor in DRK College of Engineering and Technology, JNTUH, Hyderabad, Andhra Pradesh, India. He is pursuing Ph.D in Information Security. He has completed M.Tech (C.S.E) from JNTUH. His main research interest includes Information Security and Computer Ad-Hoc Networks.



Dr.R.V.Krishnaiah, did M.Tech (EIE) from NIT Waranagal, MTech(CSE) from JNTU, ,Ph.D, from JNTU Ananthapur, He has memberships in professional bodies MIE, MIETE, MISTE. He is working as Principal in DRK Institute of Science and Technology, Hyderabad. His main research interests include Image Processing, Security systems, Sensors, Intelligent Systems, Computer networks, Data mining, Software Engineering, network protection and security control. He has various publications and presentations in various national and international journals.