Dimensionality Reduction for Dimension-specific Search

Zi Huang, Heng Tao Shen,
Xiaofang Zhou
School of ITEE
University of Queensland, Australia
{huang,shenht,zxf}@itee.uq.edu.au

Dawei Song, Stefan Rüger
Knowledge Media Institute
The Open University, United Kingdom
{d.song,s.rueger}@open.ac.uk

ABSTRACT

Dimensionality reduction plays an important role in efficient similarity search, which is often based on k-nearest neighbor (k-NN) queries over a high-dimensional feature space. In this paper, we introduce a novel type of k-NN query, namely conditional k-NN (ck-NN), which considers dimension-specific constraint in addition to the inter-point distances. However, existing dimensionality reduction methods are not applicable to this new type of queries. We propose a novel Mean-Std(standard deviation) guided Dimensionality Reduction (MSDR) to support a pruning based efficient ck-NN query processing strategy. Our preliminary experimental results on 3D protein structure data demonstrate that the MSDR method is promising.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.2.8 [**Database Applications**]: Scientific databases

General Terms: Algorithms

1. INTRODUCTION

The k nearest neighbor (k-NN) similarity search plays an central role in a wide range of applications, such as multimedia retrieval, molecular biology, medical imaging. Data objects are represented by automatically extracted content features which are points (vectors) in a high dimensional space. Similarity query processing is to find the data objects similar to a query, often the nearest k neighboring points of the query object in the high dimensional space, by measuring distance (often Euclidean distance) between each point in the database and the query point.

Conditional k-NN. Recently, there has been an emerging demand of a new type of queries, which take into account not only inter-point distances (as in the conventional $k-NN$), but also certain local constraints that two objects must match within certain tolerance threshold along certain individual dimensions. This is meaningful for many practical applications. For example, in protein structure analysis, two structures which are close in terms of their Euclidean distance but have a very large difference along certain single dimension may potentially lead to completely different biological functions [6]. We refer such type of query to as *conditional k-NN query* (ck-NN query).

Curse of dimensionality. Similarity search over a large scale complex data repository is computationally intensive, mainly due to the size of data and the high dimensionality of the feature space, known as the "curse of dimensionality" [2]. To alleviate the problem, dimensionality reduction techniques have been explored for finding a lower dimensional approximation of the original space. However, existing dimensionality reduction methods, such as Principle Component Analysis (PCA), Singular Value Decomposition (SVD) and Locality Preserving Projections (LPP)[3], are not applicable to ck-NN since they consider the global correlation of dimensions among all points and neglect the difference along each individual dimension between two points. Consequently, they are not sufficient to support ck-NN queries.

In this paper, we formulate the conditional k-NN problem and propose a novel *Mean-Std guided Dimensionality Reduction* (MSDR) to facilitate an efficient ck-NN query processing strategy.

2. CONDITIONAL K-NN SEARCH

Definition1:Dimension-specific Similarity Measure *Given two objects represented by their feature vectors* $A = (a_1, a_2, ..., a_D)$ *and* $B = (b_1, b_2, ..., b_D)$. *A and B are similar (denoted as* $A \cong_\varepsilon B$*), if* $\forall i \in 1..D$, $a_i \cong_\varepsilon b_i$*), where D is the dimensionality of the feature space.*

"$\cong_\varepsilon$" means "equal to within a tolerance ε", where $\varepsilon = 0$ implies a rigid matching (i.e., the two objects are identical); otherwise a semi-rigid matching. However, one potential problem is that, depending on the value of ε, there can be too many similar objects found. It is often necessary to add a *global similarity measure*, normally based on the Euclidian distance between objects.

In summary, given a query object, the problem we investigate here is to find the k *most similar objects* from the dataset. The similarity between two objects is measured by the Euclidean distance, subject to the maximum allowable variation on each dimension.

3. MEAN-STD GUIDED DIMENSIONALITY REDUCTION (MSDR)

MSDR first analyzes the statistics of each dimension globally. Those dimensions having the similar statistics are then summarized into single ones. For each point, each of its summarized dimension is in turn represented by the local statistic of its original dimensions. The statistical parameters we use are *mean* (μ) and standard deviation (σ). We outline the algorithm of MSDR as follows.

Compute μ and σ: MSDR first computes μ and σ of val-

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

ues alone each dimension for all points and represents each dimension as a $2D$ point (μ,σ) . The original space $\mathbb{X}$ is then transformed to a $2D$ μ-σ space with D points corresponding to the original D dimensions.

Clustering: The K-means algorithm is employed in μ-σ space to classify the D points into D' clusters ($D' = K$), each of which is described by its centroid c.

Generate the subspace: It is natural to combine the dimensions in each cluster into one, resulting in a new D'-dimensional space $\mathbb{X}'$. Each object in $\mathbb{X}$ is mapped onto $\mathbb{X}'$ where its value on each new dimension is the mean value of the dimensions in $\mathbb{X}$ that form the new one.

4. *ck*-NN QUERY PROCESSING

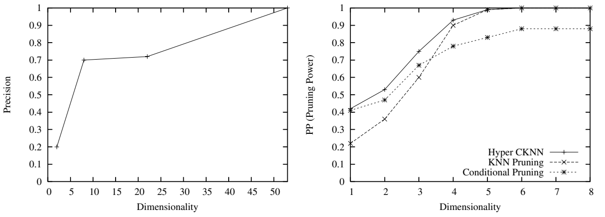
An intuitive way for *ck*-NN query processing is to prune the data space by first performing conditional pruning along individual dimensions to form a candidate set, and then further prune the candidate set by the global similarities, or vice versa. However, it will generate a relatively large number of intermediate candidates which is strongly undesirable for large scale data.

We propose a novel hybrid *ck*-NN query processing strategy which cooperates conditional pruning and k-NN pruning concurrently. The basic index structure is to maintain a separate table (objects-values) for each dimension. The tables are accessed one by one in certain order. After a dimension is accessed, the lower and upper bounds of candidates are updated correspondingly. The candidates are then sorted by their upper bounds in an ascending order. For each candidate, it is first checked by the condition rule and removed from the candidate set if the condition is violated, followed by k-NN pruning. If the current candidate's lower bound is greater than the k^{th} largest upper bound, it can be safely pruned. In the next iteration, the same process is performed and the candidate set is further reduced, until all dimensions have been processed since the measure of similar patches requires pair-wise comparisons on all dimensions. Due to the page limit, we will not give the formal description of upper/lower bound in this paper.

5. EXPERIMENTS

This section reports our preliminary performance study on our MSDR algorithm. As mentioned previously, the existing dimensionality reduction methods[3] and high-dimensional indexing methods [2] are not applicable to the *ck*-NN problem, and thus not directly comparable with the proposed MSDR approach. MSDR derives a lower dimensional space $\mathbb{X}'$ and therefore will lose information. Average precision is used as the effectiveness indicator, with *ck*-NN search results from the original space as ground truth. The efficiency of our hybrid *ck*-NN algorithm is measured by the Pruning Power (PP) which is defined as the ratio of the number of pruned objects to the total number of objects. Clearly, a larger PP corresponds to a more powerful pruning strategy, hence a faster response.

We conduct experiments on 3D protein structure data. The feature space is constructed based on a compact data representation model, which represent each structure as a high dimensional point[5][4]. From a total number of 1,100 sample protein structures in the Protein Data Bank [1], we build a dataset of 2,207,018 53-dimensional points (feature vectors). 100 points are randomly selected as queries. k is set to 10. The distance tolerance ε along each dimension is set to be $1.5\AA$ to be biologically meaningful.



(a) dim vs. precision (b) PP Comparison

Figure 1: Effect of δ and D$'$.

Fig.1(a) shows the average precision increases as more dimensions are retained. When D$'$ reaches 8, the precision increases sharply to more than 70% while the new space is only 8/53 of the original space. When D$'$ is greater than 8, the precision then increases slowly before finally reaching 100%.The result confirms that MSDR can achieve reasonable quality of results even when the dimensionality is reduced to be a very small number. Fig.1(b) depicts the pruning power comparison of three methods, from which we can observe that our hybrid *ck*-NN method achieves the best performance which prunes more space than the rest two methods by large percentages.

6. CONCLUSION

In this paper, we propose a new query type called *ck*-NN query and correspondingly a novel method MSDR to support efficient *ck*-NN search. The results demonstrate an encouraging performance of our method. More extensive performance evaluation is currently being conducted on protein structure as well as image data and will further involve comparison with human relevance judgments.

7. ACKNOWLEDGEMENTS

This work was funded in part by the UK's Engineering and Physical Sciences Research Council (EPSRC) through an Overseas Travel Grant (EP/E037402/1), and the European Union Sixth Framework Programme (FP6) through the integrated project Pharos (IST-2006-045035).

8. REFERENCES

[1] Protein data bank. http://www.rcsb.org/pdb/.

[2] C. Böhm, S. Berchtold, and D. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, pages 33(3):322–373, 2001.

[3] X. He, D. Cai, H. Liu, and W.Y. Ma. Locality preserving indexing for document representation. In *SIGIR*, pages 96–103, 2004.

[4] Z. Huang, X. Zhou, H.T. Shen, and D. Song. 3d protein structure matching by patch signatures. In *DEXA*, pages 528–537, 2006.

[5] Z. Huang, X. Zhou, D. Song, and P. Bruza. Dimensionality Reduction in Motif-Signature Based Protein Structure Matching. In *ADC*, pages 89–97, 2006.

[6] R.V. Spriggs, P.J. Artymiuk, and P. Willett. Searching for patterns of amino acids in 3D protein structures. *J Chem Inf Comput Sci.*, 43(2):412–421, 2003.