IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 06-10
www.iosrjournals.org

Nearest neighbor search with anytime clustering method

Miss. P. Y. Mahurkar,.pragatimahurkar@gmail.com
Dr. R. D. Raut, SGBAU, Amravati, India ,drrautrd@gmail.com
Dr. V. M. Thakare, SGBAU, Amravati, India ,vilthakare@yahoo.com

Abstract: *The rapidly increasing volume of multimedia content over the Internet creates many research challenges for efficiently storing, processing, and searching such sheer volume.In general, most algorithms need two types of data structures named index data and feature data, both of which are frequently visited during k-NN search. Object matching is an important problem with numerous real-life applications, such as document classification. To solve this problem, a method for finding the nearest neighbors of a given data point is needed and mainly it depends upon clustering strategy. This paper presents the techniques such as DIMO, a distributedindex system for matching high-dimensional multimedia objects, multiplicativedistance function, k-means, ONE (Online Nearest-neighbor Estimation),k-nearest neighbor (kNN), CKNN (Continuous k nearest neighbor). To increase the speed and efficiency of finding the similar objects and to maintain the flexibility of the dynamic nature of the objects, this paper propose a novel Anytime Clustering method to find the similar objects.This method can give result efficiently.*

Keywords :-*DIMO, multiplicativedistance function, k-means, ONE, kNN, CKNN, Clustering Anytime*

I. INTRODUCTION

K Nearest neighbor search is one frequently used category of algorithms for solving similarity search problem. There is a concept “feature” to represent one data item in the database. In general, a feature *f* can be defined as a *D* dimensional vector. The database*X* is defined as a set of *N* such features: $X = \{f1, f2, . . . , fN\}$. The similarity is often measured by Euclidean Distance (ED). Based on these definitions, the *k*-NN problem can be formally described as: given a query feature *q*, find *k* reference features in *X* that have the shortest (Euclidean) distances to *q*. There are two approaches which are computes exact neighbors and approximate neighbors.

There is a space partitioning method in which the data are divided to some partitions and the nearest neighbor or distanceis obtained over these partitions. These methods have some drawbacks with increasing the dimensionality like significant increase in the number of partitions and complexity and the performance degradation. Moreover, some of the methods are just applicable for the nearest neighbor searchand cannot give the exact value of distance. The index structure is used for finding reference Features called candidate features that are most likely to be the *k* nearest neighbors. To decide whether a candidate feature is one of the *k* nearest neighbors, the feature data is visited in order to evaluate their similarity. This paper discusses methods DIMO structure, multiplicativedistance function, K-meansONE (Online Nearest-neighbor Estimation), *k*-nearest neighbor (kNN), CKNN. To improve the speed and performance of the algorithm, clustering at anytime nature is adopted.

II. BACKGROUND

A C*k*NN query is a query that continuously finds the *k* nearest neighbors with respect to a given moving point-object among a set of *n* moving point-objects. In general, C*k*NN queries are used in any database application that continuously ranks answers based on their distances to an object.There is an uncertainty ofobject locations, the set of actual k nearest neighbors isunknowable.The CKNN algorithm partitions the data objects into two approximately equal-sized groups and recursively maintains the minimum of each group[1].DIMO provides multimedia applications with the basic function of computing the K nearest neighbors on large-scale datasets. DIMO presents a novel method for partitioning, searching, and storing high-dimensional datasets on distributed infrastructures that support the MapReduce programming model.DIMO is designed for large-scale datasets. DIMO can automatically use varying number of computing machines. Data points can be added and removed from the system in dynamic manner, without the need to re-build the whole system [2].In multiplicative distance function, there is stability of data with independent dimensions with identical or non identical distribution. It has been shown that the commonlyused distance functions become unstable in high-dimensional data space; i.e., the distanceto the farthest data point approaches the distance to the nearest data point of a givenquery point with increasing dimensionality.In contrast to usual distancefunctions which are based on the summation of distances over all dimensions, the multiplicative distance is based on the multiplication of distance components [3].In ONE method, there is an online estimation of nearest neighbors. It is a unified algorithm for both image classification and retrieval using text. This is achieved by computing similarity between the query and each category or image candidate. ONE is surprisingly simple, which only involves manual object definition, regional description and nearest-neighbor search. ONE achieves state-of-the-art accuracy in a wide range of image classification and retrieval benchmarks [4]. The*k*-nearest neighbor (*kNN*) is one of the most simple and widely used methods. Itis based on a relaxed definition of the local feature based image to image similarity and allows standard *kNN*classification to be efficiently executed with the support of access methods for similarity search. Given a query image, the *kNN*algorithm scans a training set to retrieve the best-matching images [5].

This paper introduced the efficient methods for nearest neighbor matching i.e. **Section I.** Introduction. **Section II.**discuss Background. **Section III.**discuss Previous work done. **Section IV.**discuss Existing methodology. **Section V.** Analysis And Discussion. **Section VI.**Proposed methodology.**Section VII.** Outcomes and possible result.**Section VIII.** Conclude paper finally **Section IX.** Future scope.

III. PREVIOUS WORK DONE

A. prasadSistla et al [1] (2014) has worked on a C*k*NN query is a query that continuously finds the *k* nearest neighbors with respect to a given moving point-object *Oq* among a set of *n* moving point-objects in a certain case. The query returns a sequence of answer-pairs, namely pairs of the form *(I, S)* such thatis a time interval and *S* is the set of *k* objects that are nearest to *Oq*during *I*. the answer pairs are required to be non-redundant.While the uncertainty model is a circular region and the object is anywhere in this region. This captures the uncertaintycaused by positioning errors. It is also a result of the reduction of the communication/energy cost for a movingobject to report its location changes. In this policy, a movingobject sends an update to the database when the distancebetween its current location and the location stored in thedatabase exceeds a certain threshold. Mohamed Hefeeda et al [2] (2014) has worked on DIMO which DIMO takes two sets: reference points R and query points Q. Each set contains d-dimensional data points. There are no constraints on the sizes of Q and R. However, R is assumed to change at a slower rate, by adding/removing objects to/from it. Whereas Q can change faster. Distributed index is divided into two parts directing tree and bins.Directing tree is used to group similar points in the same or close-by bins as well as to forward query points to the bins with potential matches. Bins are the leaf nodes of the directing tree, but they are stored as files on the distributed file system. The distributed index is constructed from the reference R dataset, which is done before processing any queries. Data points are stored only at the leaf nodes. JafarMansouri et al [3] (2015) has worked onthe multiplicative distance which contains the production of elements. The multiplicative distance can resist to instabilityin high-dimensional space for a wide range of data distributions. There is a case in which calculating the distance causes overflow. To remove this problem, the small *c* can be used. The overflow depends on the programming language, number of dimensions, and distribution of data. So, usually, the value of *c* cannot be determined in advance. The performance of multiplicative distance function is measured by the real application such as clustering algorithm k-means.Qi Tian et al [4] (2014) has worked on online nearest neighbor estimation technique which allows classification by using text. Image retrieval is closely related to a number of real-world multimedia applications. Given an image database and a query, it requires finding relative candidates to the query in a short time.On the online querying stage, the query's relevance to a category or candidate image is estimated by the averaged nearest distance from querying objects to the objects in that category or candidate image. Nearest neighbor classification works as follows: first create a database of objects, then the system is given a query then system simply finds the nearest neighbor of the query in the database. Giuseppe Amato et a [5] (2015) has worked on KNN method which scans a training set to retrieve the best-matching images for a given query. The execution of the *kNN*classification algorithm requires thatthe query image be sequentially compared with all images of the training set. The *kNN*classification allows to classify images and relies on a relaxed definition of the local feature based image to image similarity definition, which allows efficient index for similarity search to be used. A kNN performssearch between the objects of the training set *T*. it contains the documents with labels. After processing, the documents are classified.

The proposed methodis focused on a nearest neighbor matching by anytime clustering approach. It is a novel and flexible method to mitigate problem. This interactive methodology can find nearest neighbor match very fast. It allows a user to achieve high performance for the desired data. When dealing with a large datasets, this method works by partitioning the dynamic objects into clusters at anytime. In future, the distributed nearest neighbor matching technique can be applied to the additional larger real data sets and higher dimensional data to enhance their robustness and quality. Experiments show that proposed system can handle a clustering of a large variety of input data. For future studies, researchers are trying to explore the graphical data for clustering and similarity matching in a limited time.

IV. EXISTING METHODOGY

CKNN method: The processing of CKNN is carried in two approaches online processing and offline processing. For off-line processing in the certain case, there is simple divide and conquer algorithm which gives the solution and database is disk based or main memory based. For on-line processing in the certain case, there is a kinetic data structure, called object heap. This data structure allows updates like insertion of a new object, deletion of an existing object and velocity-vector change of an existing object [1].

DIMO method: DIMO partitions the reference points R into bins. These bins are mapped to files and stored on a distributed file system. The bins are searched in parallel against query objects.The core component in our system that enables efficient partitioning, mapping, and searching for objects is the Distributed Index. There are two main computational tasks: Build Index and Match Objects. Build Index takes the reference points and creates the Distributed Index. Match Objects computes the nearest neighbors for each query point as well as it performs application-specific object matching functions using the found nearest neighbors [2].

*Multiplicative distance method:*The multiplicative distance function is applied on k-means algorithm to search the nearest neighbors. In this, the objects are partitioned into k clusters in which the query object is matched. Clusters have different shapes. Multiplicative distance is a new distance function which is theoretically proved for data with independent dimensions, its Pearson variation does not tend to zero when dimensionality increases to infinity. Simulation results show the stability of this distance function for correlated data, too. As an application, it is shown that this distance function has better performance than the norm distances for clustering [3].

*ONE method:*In ONE technique, Bothclassification and retrieval involve measuring the similarity between the query and training or candidate images. The only difference lies in that a class label is provided for each training sample in classification. Therefore, it is possible to actually computing the similarity between the query and a category, i.e., a set of images [4].

*KNN method:*KNN is used for image classification and retrieval. The idea of applying the *kNN*classification in combination with the geometric consistency technique is very effective for tasks where only a few objects need to be recognized and the training sets are small. The execution of the *kNN*classification algorithm requires that the query image be sequentially compared with all images of the training set. To compare the query image with a single image of the training set, all local features of the query image must be compared with all local features of the training set image [5].

*DATASET:*The performance of the DIMO system is assessed using data points extracted from images. Two image datasets used in experiments: Caltech Dataset. This dataset is composed of data points extracted from 2,500 images. The images are obtained from the Caltech Buildings and Game Covers datasets. ImageNet Dataset. ImageNet is an open image database with millions of images organized according to the WordNet hierarchy. For CKNN, The test data were provided by Shanghai Jiaotong University.The data contain GPS traces collected from over 4,000 taxis running in the Shanghai urban area for 28 days.

V. ANALYSIS AND DISCUSSION

In CKNN, it can be seen that when there are fewer than 30 updates, off-line is better than on-line. However, when there are many updates, on-line is much better than off-line. Also, the uncertain case has much more answer-pairs than the certain case. Without any indexing structure, query processing easily scales to 1 million objects in the certain case and it scales to 10,000 objects in the uncertain case but it requires very large space [1]. In DIMO, index is general and can be used by multiple applications that require nearest neighbors search in high-dimensional spaces. For insertion and deletion of objects, the whole tree needs to be traversed to locate on the bins. It has time complexity on the large set of data. It requires a high maintenance cost to manage the index, bins [2]. The performance of multiplicative function is better than additive function. Themultiplicative distance function suffering from some instability problems. The k-means requires a large searching time and complexity increases as the data size increases [3]. ONE achieves accuracy in the classification works. However, it is worth noting that the actual computational costs in ONE are much more expensive than conventional algorithms [4]. KNN technique performed efficiently and effectively. But it is very costly and accuracy of the KNN is degraded by the presence of noisy and irrelevant features [5].

Querysearch techniques	Advantages	Disadvantages
CKNN **Query search**	1) In this method, there is no need of index structure. 2) It has efficient processing.	1) It is consumes lots of space. 2) There is an uncertainty due to frequently updation.
DIMO **method**	1) It requires less memory for processing. 2)It balances the load across the used computing machines.	1) It has a time complexity for a very large data. 2) Insertion and deletion cause high maintenance cost.
k-means **search**	1) It is efficient to run a k-means cluster. 2) It is effectivefor large dataset.	1) It lacks consistency. 2) Its working is depend on the shape of clusters.
ONE **technique**	1)Accuracy is there in the classification. 2)it requires less storage.	1) Large and complex cases do not give suitable answer. 2) It requires longer time.
KNN **search**	1) Larger values of k reduce the effect of noise. 2) It gives consistent output.	1)It is sensitive to the local structure of data. 2)Severely degraded by the presence of noise.

Table 1: Comparison between different nearest neighbor searching techniques

VI. PROPOSED METHODOLOGY:

Anytime Clustering Method

The proposed method anytime clustering allows the task of assigning objects into clusters such that the similarity of objects within a group is maximized and the similarity of objects between different groups is minimized. For the online processing of queries, objects are continuously moving. So it is better way to partition the objects according to their similarity. Clustering is nothing but the partitioning of similar contents into the same group.

Fig. 1 shows the overall architecture that reads the data, check the similarity between it and cluster the data. Then it becomes efficient to match the nearest neighbor. As it takes less time to search the similar data hence, performance is high with this method.

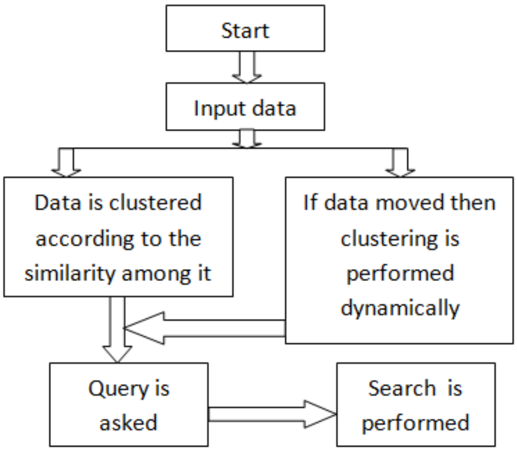
Start
Input data
Data is clustered according to the similarity among it
If data moved then clustering is performed dynamically
Query is asked
Search is performed

Fig. 1: Anytime Clustering technique

The proposed method works as follows:
1. Analyze data and cluster it **:**The data is analyzed and partitioned into clusters according to the similarity between it.
*2. Nearest neighbor search***:** The query is fired and search is performed on the clusters of the related type.
3. Again cluster if location changed:Although the location of data is changed, the clustering is performed dynamically within a short time and again search is performed.

VII. OUTCOME AND POSSIBLE RESULT

The result of this method focused on clustering the data efficiently and dynamically. The proposed method performed well and effectively as compared to the existing methods.The purpose here is to cluster data at any time and complete the query search. It consumes less storage space and requires less iteration because of clustering. This method is the most accurate one.

VIII. CONCLUSION

Thus, this method focused on partition the dataset into clusters based on the similarity among data and if the location is changed then again perform dynamic clustering and query search is completed with the help of clusters. The performance of this method is flexible as compared to the other techniques. The output of this method provide best result as it gives certainty in the result and can handle a large data.

IX. FUTURE SCOPE

In future, this method can be applied on the graphical data and very high dimensional data for query searching. Researchers are trying to explore the graphical data for clustering and similarity matching in a limited time.

REFERENCES

[1]. A. Prasad Sistla, OuriWolfson and Bo Xu, “Continuous nearest-neighbor queries with location uncertainty”, Springer-Verlag Berlin Heidelberg,Vol no 24, PP 25-50, 10 June 2014

[2]. Ahmed Abdelsadek and Mohamed Hefeeda, “DIMO: Distributed Index for Matching Multimedia Objects using MapReduce”, MMSys ’14, March 19–21 2014, Singapore Copyright,Vol no 14, PP 114-126,ACM 2014

[3]. LingxiXie, Richang Hong, Bo Zhang and Qi Tian, “Image Classification and Retrieval are ONE”, ICMR’15, June 23–26, 2015, Shanghai, China. Copyright,Vol no 15, PP 3-10, ACM 2015

[4]. JafarMansouri and MortezaKhademi, “Multiplicative distance: a method to alleviate distance instability for high-dimensional data”, © Springer-Verlag London, Vol no 25,PP 783-805, 28 December 2014

[5]. Giuseppe Amato, FabrizioFalchi and Claudio Gennaro, ISTI-CNR, “Fast Image Classification for Monument Recognition”, Giuseppe Amato, FabrizioFalchi, and Claudio Gennaro. 2015. Fast image classification for monument recognition. ACM J. Comput. Cult.Herit. 8, 4, Article 18, PP 1-25, August 2015.