International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 31-35

Parallel and Nearest Neighbor Search for High-Dimensional Index Structure of Content-Based Data Using Dva-Tree

A.RosiyaSusaiMary[1], Dr.R.Suguna [2]
[1](Department of Computer Science, Theivanai Ammal College for Women, India)
[2](Department of Computer Science, Theivanai Ammal College for Women, India)

Abstract: *We propose a parallel high-dimensional index structure for content-based information retrieval so as to cope with the linear decrease in retrieval performance. In addition, we devise data insertion, range query and k-NN query processing algorithms which are suitable for a cluster-based parallel architecture. Finally, we show that our parallel index structure achieves good retrieval performance in proportion to the number of servers in the cluster-based architecture and it outperforms a parallel version of the VA-File. To address the demanding search needs caused by large-scale image collections, speeding up the search by using distributed index structures, and a Vector Approximation-file (DVA-file) in parallel.*

Keywords: *Distributed Indexing Structure, High Dimensionality, KNN- Search*

I. Introduction

The need to manage various types of large scale data $_{\text{stored}}$ in web environments has drastically increased and resulted in the development of index mechanism for high dimensional feature vector data about such a kinds of multimedia data. Recent search engine for the multimedia data in web location may collect billions of images, text and video data, which makes the performance bottleneck to get a suitable web documents and contents. Given large image and video data collections, a basic problem is to find objects that cover given information need. Due to the huge amount of data, keyword based techniques are too expensive, requiring too much manual intervention. In contrast, a content-based information retrieval (CBIR) system identifies the images most similar to a given query image or video clip.

II. Related Works

There is an underlying need of indexing techniques that should support execution of similarity queries. A web search engine that provides the contents-based retrieval upon the multimedia data requires complex distance functions to measure similarities of multi-dimensional features, to search a data based on the user query.

This author signature based [1] new efficient high dimensional indexing scheme to support the content-based retrieval of a large amount of video data. For this, we extend Hybrid Spill-Tree by using two techniques, such as a signature technique and a newly designed clustering technique. By adopting a signature technique, we can reduce both storage overhead and the retrieval time. By using a new sampling technique, we can greatly reduce the computational cost in case of inserting high-dimensional vector data. In addition, we provide two algorithms for our efficient high dimensional indexing scheme; an insertion algorithm and a k-NN search algorithm. [2] The content based implemented the clustering techniques are discussed and analyzed. Also, we propose a method HDK that uses more than one clustering technique to improve the performance of CBIR. These methods makes use of hierarchical and divide and conquer K-Means clustering technique with equivalency and compatible relation concepts to improve the performance of the K-Means for using in high dimensional datasets. It also introduced the feature like color, texture and shape for accurate and effective retrieval system. [3] Visual content based video indexing presented a review on recent developments in visual content-based video indexing and retrieval. The state of the art of existing approaches in each major issue has been described with the focus on the following tasks: video structure analysis including shot boundary detection, key frame extraction and scene segmentation, extraction of features of static key frames, objects and motions, video data mining, video classification and annotation, video search including interface, similarity measure and relevance feedback, and video summarization and browsing. [4] This Author proposed a novel algorithm, namely, Cluster-based Temporal Mobile Sequential Pattern Mine (CTMSP-Mine), to discover the Cluster-based Temporal Mobile Sequential Patterns (CTMSPs). Moreover, a prediction strategy is proposed to predict the subsequent mobile behaviors. In CTMSP-Mine, user clusters are constructed by a novel algorithm named Cluster-Object-based Smart Cluster Affinity Search Technique (CO-Smart-CAST) and similarities between users are evaluated by the proposed measure, Location-Based Service Alignment (LBS-Alignment). Meanwhile, a time segmentation approach is presented to find segmenting time intervals where similar mobile characteristics exist. To our best knowledge, this is the first work on mining and prediction of mobile behaviors with considerations of user relations and temporal property simultaneously. Through experimental evaluation under various simulated conditions, the proposed methods are shown to deliver excellent performance. [5] Multi biometric method for generating fixed-length codes for indexing biometric databases. An index code is constructed by computing match scores between a biometric image and a fixed set of reference images. Candidate identities are retrieved based on the similarity between the index code of the probe image and those of the identities in the data-base. The proposed technique can be easily extended to retrieve pertinent identities from multimodal databases. Experiments on a chimeric face and fingerprint bimodal database resulted in an 84%average reduction in the search space at a hit rate of 100%. These results suggest that the proposed indexing scheme has the potential to substantially reduce the response time without compromising the accuracy of identification. Shape based image [6] clustering a novel approach is proposed for medical image retrieval based on shape feature, which uses canny edge detection algorithm for extraction of image shape and K-means algorithm for extraction of different regions of the image in order to improve better matching process between user query image and feature database images. We have shown that Canny Edge Detection and K-means clustering algorithms are quite useful for retrieval of relevant images from image database. Our results indicate that the proposed approach offers significant performance improvements in retrieval of medical images. [7] This authorproposed Despite the considerable progress of academic research in video retrieval, there has been relatively little impact of content based video retrieval research on commercial applications with some niche exceptions such as video segmentation. Choosing features that reflect real human interest remains an open issue. One promising approach is to use Meta learning to automatically select or combine appropriate features. Another possibility is to develop an interactive user interface based on visually interpreting the data using a selected measure to assist the selection process. Extensive experiments comparing the results of features with actual human interest could be used as another method of analysis. [8]content based various relevance feedback techniques for last ten years, their dataset used and their results are discussed in detail. From the results of the various methods discussed, it can be concluded that to improve the retrieval performance of the CBIR systems researchers must have to design the techniques to increase the values of the standard evaluation parameters like precision, convergence ratio or accuracy. The Relevance Feedback technique can be incorporated in CBIR system to obtain the higher values of the standard evaluation parameters used for evaluation of the CBIR system which may lead to better results of retrieval performance. For future research direction in RF, the approaches discussed can be applied to more kinds of applications on multimedia retrieval or multimedia recommendation.

III. Overall Procedure

The high dimensional data space is transformed into a Distributed vector approximation tree in our method. In this construction step, the tree just considers sampled feature data from the original high dimensional data space. The leaf node which should be accommodated in a distinct machine has a local signature file for the correspondingrange of original feature vector. Thus we can support the parallel search through the leaf nodes and ensure the efficient response for the contents-based retrieval searching by fig.1. The index should be deployable over multiple nodes in cluster environments. The index should require no special tuning of parameters required for each specific dataset. The set of candidates retrieved by the index should contain the most similar objects to the query. The fig (4) represents the process of proposed system. Using the diagram algorithm also explained in the form of tree index hashing.

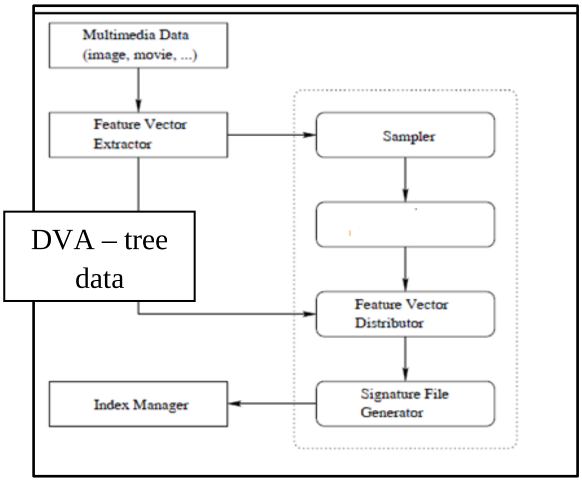

Fig 1: Data searching

One Day National Conference On "Internet Of Things - The Current Trend In Connected World" NCIOT-2018 32 | Page

The number of candidates retrieved must be as small as possible, to reduce I/O and computation costs. On the other hand, most multi-dimensional indexing structures have an exponential dependence upon the number of dimensions. In recognition of this, a VA-file [9] was developed to accelerate the scan through the feature vectors. The VA-file consists of two separated files: the vector file containing the feature vectors, and the approximation file containing acompressed representation of each feature vector (fig.2).

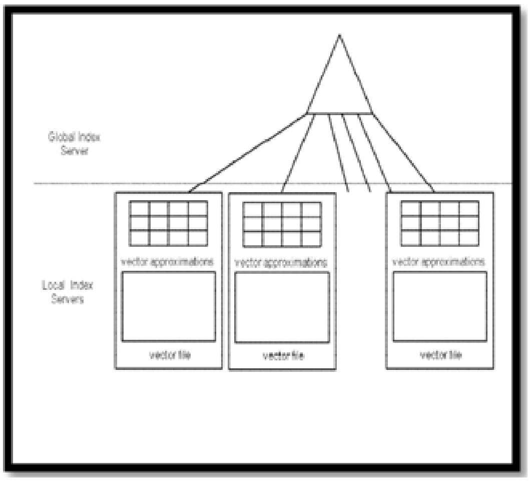

Fig 2: Structure of the DVA-tree

IV. Experimental Results

The performance is evaluated using the average execution time and accuracy of a k-NN search over 10 different queries. We compare the Performance of the DVA-tree with that of the distributed hybrid spill-tree fig.4 is used because the distributed hybrid spill-tree is a recent indexing structure based on a cluster environment. The distributed hybrid spill-tree and DVA-tree algorithms were developed using the M-tree C++ package. This is to construct a similar query execution environment as the Map Reduce operations for the nearest neighbor search in order to emulate a larger configuration of data server. For a fair performance comparison, the top trees of the DVA-tree built on same sample data, and all the indexing structures have the same number of index servers. The DVA-tree outperforms the distributed content based-tree in terms of execution time as the number of data increases (fig.5).

This is due to the fact that local index servers in the DVA-tree utilize the VA-file technique without any processing overhead of the directory of the tree.

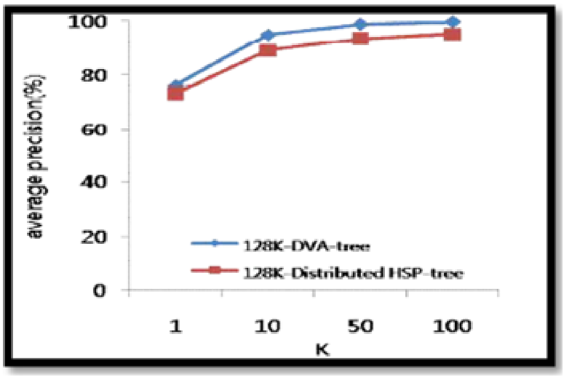

Fig 3: Data uploading speed

4.1 High dimensional index structure

Most of the distributed high-dimensional indexing structures provide a reasonable search performance especially when the dataset is uniformly distributed. However, in case when the dataset is clustered or skewed, the search performances gradually degrade as compared with the uniformly distributed dataset.

A method of improving the k-nearest neighbor search performance for the distributed vector approximation-tree based on the strongly clustered or skewed dataset. The basic idea is to compute volumes of the leaf nodes on the top-tree of a distributed vector approximation-tree and to assign different number of bits to them in order to assure an identification performance of vector approximation. In other words, it can be done by assigning more bits to the high-density clusters. We conducted experiments to compare the search performance with the distributed hybrid spill-tree and distributed vector approximation-tree by using the synthetic and real data sets. The experimental results of our proposed scheme provides consistent results with significant performance improvements of the distributed vector approximation-tree for strongly clustered or skewed datasets.

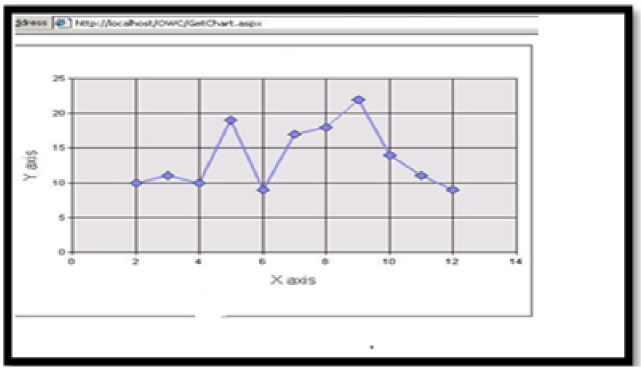

Fig 4: Searching time in dataset.

4.2 K-D-B Algorithms:

K-D-B-tree is to provide the search efficiency of a balanced k-d tree, while providing the block-oriented storage of a B-tree for optimizing external memory accesses. K-D-B-tree may require the splitting of a page in the case of a page overflow, it is important to first define the splitting operation. The structure of the K-D-B treecontains two types of pages:

Region pages: A collection of *(region, child)* pairs containing a description of the bounding region along with a pointer to the child page corresponding to that region.

Point pages: A collection of *(point, location)* pairs. In the case of databases, *location* may point to the index of the database record, while for points in *k*-dimensional space, it can be seen as the points coordinates in that space.

Page overflows occur when inserting an element into K-D-B-tree results in the size of a node exceeding its optimal size. Since the purpose of the K-D-B-tree is to optimize external memory accesses like those from ahard-disk, a page is considered to have overflowed or be overfilled if the size of the node exceeds the external memory page size.

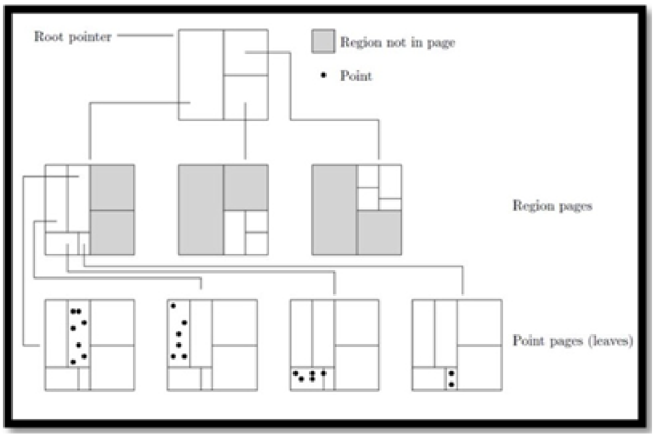

Fig 5: The basic structure of K-D-B Tree

K-D-B-Tree algorithm used to searching the side to be divided (dimension of subdivision), and the calculation of the subdivision value according to this dimension.

The Dimension of subdivision and the Value of subdivision are defined as follows:

- Dimension of subdivision: the subdivision is carried out according to the dimension having the maximum of data dissemination, i.e. the dimension that corresponds to the greatest difference between the vectors components.
- Value of subdivision: it is the value of quantification nearest to the median value according to the dimension of subdivision.

4.3 K-N-N Algorithm

Conventional index structures provide various nearest-neighbor search algorithms for high-dimensional data, there are additional requirements to increase search performances, as well as to support index scalability for large-scale datasets. To support these requirements, in this paper we

Distributed high-dimensional index structure based on cluster systems, called a Distributed Vector Approximation-tree (DVA-tree), which is a two-level structure consisting of a hybrid spill-tree and Vector Approximation files (VA-files).

$$L_2(P,Q) = d(P,Q) = \sqrt[2]{\sum_{i=0}^{d-1} (Q_i - P_i)^2}$$

…….. (1)

4.3.1. Definition 1 Range Query

Given a query object Q∈D and a maximum search distance r, the range query range (Q, r) selects all indexed objects Oj such that d (Oj, Q) ≤ r. These methods can prune the search space for queries using the partitioning.

4.3.2. Definition 2 k- K-D-B-Tree

Given a query object Q∈D and an integer k≥1, the k-NN query NN (Q, k) selects the k indexed objects which have the shortest distance from Q.

$$d_{\min}(T(O_r)) = \max\{d(O_r, Q) - r, 0\}$$

……….. (2)

$$d_{\max}(T(O_r)) = d(O_r, Q) + r$$

………….. (3)

V. Conclusion

The design of a new high dimensional indexing scheme, called a DVA-tree, to solve the distributed k-nearest neighbor search problem over large scale high-dimensional data in cluster environments. The DVA-tree employs a hierarchical clustering method and distributed VA-file management in order to allow a parallel KNN search on each of the VA-files. The sample data of large-scale high-dimensional data, because the sampling is independent of the dimensionality and the sampled data maintain the cluster information of the data set stored in the database.

References

[1]. Hyun-Jo Lee, & Jae-Woo Chang, “ Signature-based Hybrid Spill-Tree for Indexing High-Dimensional Data”, *IEEE Ninth International Conference on Computer and Information Technology, 978-0-7695-3836-5/09 $26.00 © 2009 IEEE DOI 10.1109/CIT.2009.93.*

[2]. Mrs. Monika Jain, & Dr. S.K. Singh, “A Survey On: Content Based Image Retrieval Systems Using Clustering Techniques for Large Data sets”, *International Journal of Managing Information Technology (IJMIT) Vol.3, No.4, November 2011`*

[3]. Weiming Hu, NianhuaXie, Li Li, XianglinZeng, and Stephen Maybank,” A Survey on Visual Content-Based Video Indexing and Retrieval”, *IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 41, no. 6, november 2011.*

[4]. Eric Hsueh-Chan Lu, Vincent S. Tseng, and Philip S. Yu,” Mining Cluster-Based Temporal Mobile Sequential Patterns in Location-Based Service Environments”, *IEEE transactions on knowledge and data engineering, vol. 23, no. 6, June 2011.*

[5]. AglikaGyaourova, and Arun Ross, “Index Codes for Multibiometric Pattern Retrieval”, *IEEE transactions on information for ensics and security, vol.7, no.2, April 2012.*

[6]. B. Ramamurthy, and K.R. Chandran,” CBMIR: shape-based image retrieval using canny edge detection and k-means clustering algorithms for medical images”, B.Ramamurthy et al. / *International Journal of Engineering Science and Technology (IJEST), ISSN: 0975-5462 Vol. 3 No. 3 March 2011.*

[7]. B V Patel, and B BMeshram,” content based video retrieval systems*”, International Journal of UbiComp (IJU), Vol.3, No.2, April 2012.*

[8]. 1Latika Pinjarkar, Manisha Sharma, and Kamal Mehta,” Comparison and Analysis of Content Based Image Retrieval Systems Based On Relevance Feedback “, *VOL. 3, NO.6, July 2012, ISSN 2079-8407, Journal of Emerging Trends in Computing and Information Sciences.*