Cluster Optimization for Similarity Process using De-Duplication

Dileep Kumar Kadali[1] Dr. R.N.V. Jagan Mohan[2] M. Srinivasa Rao[3]
[1,3]Student [2]Professor
[1,2,3]Department of Computer Engineering
[1,2,3]Swarnandhra College of Engineering & Technology, Seetharampuram, Narsapur-534280

Abstract— Over the decades, voluminous of data revolution where nearly every aspect of computer engineering is being driven by large-data processing and analysis. Valid data is important for accessing system without De-Identification is a well-known technique for recognizable system which provides the privacy for individuals. This system incorporates large data sets i.e., Big Data by forming clusters. In hierarchical clustering, the output is a tree giving a sequence of clustering, with each cluster being a partition of the dataset. The major drawbacks of the clusters in automatic research with the choice of the most relevant features are not compared to the similarity process (Duplication). However, the system suffers to how to reduce the computational load on leaf images. In this paper, an algorithm proposed on personnel signature based foliage images to optimize the automatic feature-subset selection based T-nary clusters and also classification of clustering is used for support vector machine with the help of Map Reduce Technique. The main focus is to concentrate on foliage images i.e., particularly for high-dimensional data sets and which are not at all relevant for a given operation. Also the low-dimensional feature subspaces are used to form the number of clusters and which are used to decide the cluster centers with the most relevant features at a faster pace.

Key words: Big Data, Foliage Images, De-Identification, Support Vector Optimization and Map Reduce

I. INTRODUCTION

This kind of this, personnel recognition system is useful to the privacy for individual. Suppose personnel signature is the personnel recognition system. Usually, personnel signatures uses the technique of machine recognition of personnel images, is one of the challenging areas of research for those working in the field of automatic pattern recognition and classification [8]. The personnel signatures detection is broadly classified into two ways: Detection of individual signatures, recognition of words in a language and recognition of personal signatures. The personal signature must be segmented into individual digits and then these digits must be classified and labeled before a complete recognition of the code is achieved. Segmentation is a difficult problem. Even it seems to be a simpler problem of individual user recognition, is an unsolved problem.

At this point, this consider the problem of personnel signature recognition. This problem may be approached in one of two modes namely called On-line and Off-line proposed by Jitao et al., has given the idea of Learn to Personalized Image Search from the Photo Sharing Websites, IEEE Transactions on multimedia, 2012[3]. At first, On-line recognition system make use of a special hardware to obtain dynamic information and use it for classification. However, they are not attractive for massive use because of special hardware requirement. In off-line recognition, the digits written on a conventionally used material like paper or envelope are scanned, digitized and stored on a machine. This data is used for recognizing the digits. Here, the recognition system does not have access to dynamic information of the digit like the number of signatures. So, classification using the offline data is more difficult. However, offline recognition is more popular because of its realistic viability. Proposed paper deals with offline recognition of Personalized Images.

To project is system it, incorporates a large data set i.e., Big Data by forming clusters. Clusters has played a critical role for pattern recognition, image segmentation i.e., which provides similar objects to create groups in an un-labeled data. Various clustering algorithms are created to define partitioning of a dataset. It is importance of parameters are assigned an improper value, the clustering method results in a partitioning scheme that is not optimal for the specific data set leading to wrong decisions. The problems of deciding the number of clusters better fitting a dataset as well as the evaluation of the clustering results has been subject of many research efforts. Previous works on these lines were proposed by various authors: discussed an application of DE to the automatic clustering of large unlabeled data sets in Automatic Clustering Using an Improved Differential Evolution Algorithm proposed by Swagatam Das, Ajith Abraham and Amit Konar, 2008[1].

The rest of the paper is organized as follows: Section 1 deals with present surveys of related work. Section 2 deals with mathematical based Ternary clusters are defined. Section 3 deals with support vector machine. Section 4 contains details about Map Reduce Technique. Section 5 presents the study of Similarity based Ternary Cluster Personalized Images Algorithm. Section 6 presents the experimental results obtained. Finally, this paper concludes the conclusion and future works in Section 7.

II. TERNARY CLUSTERS

Entire Input data can be maintained into three clusters namely user, images and tagging. In mathematical term ternary is three parts, Thus taking three variables U, I, T denote the sets of users, Images and Tags introduced by Jitao et al., given the idea of Learn to Personalized Image Search from the Photo Sharing Websites, IEEE Transactions on multimedia, 2012 [3]. The group of data of a cluster is designated by C⊂U X I X T, i.e., each triplet (u, i, t) Є C means that user u has annotated image I with tag t. the ternary interrelations can then constitute a three dimensional tensor y Є $R^{|U| X |I| X |T|}$ is applied on Map Reduce with support vector machine.

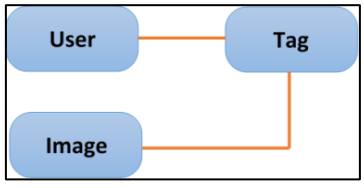

Fig. 1: Ternary Cluster Relationships

III. SUPPORT VECTOR MACHINE

In this Section, SVM is a machine learning process. It is defined over a vector space in which the problem is to find a decision surface that "best" separates the data vectors into two classes suggested by Isabelle Moulinier et al., 1997 [2]. Vipin Kumar et al., 1999 [7] is discussed the simplest linear form; an SVM is a hyper-plane that separates a set of positive from a set of negative with maximum margin.

The hyper-plane (dot) used on the database images and represented as linear SVMs or it can be found in a higher-dimensional space by transforming the images into a representation having more dimensions like input variables than the whole database in data images are treated as non-linear SVMs. It is used to provide simple solution by mapping the image data into a higher dimensional space and then reducing the problem to a linear.

The purpose of hyper-plane is to separates the training data by a maximal margin. All vectors lying on one side of the hyper-plane are labeled as 0, and all vectors lying on the other side are labeled as 1. The training instances that lie closest to the hyper-plane are called support vectors [4].

The Linear Support Vector Machine S = W.I (3.1)

Where w is the normal vector namely called as database to the hyper-plane, and i is the input vector is known as Input Data. In the linear case, the margin is defined by the distance of the hyper-plane to the nearest of the positive and negative.

IV. MAP REDUCE

Map Reduce technique is a process and used for large amount of data set by dividing the various clusters. It was introduced by Google as a new processing technique for handling large-scale data analysis. It is simplifying the distribution of application system by providing two interfaces: Map and Reduce. To build applications on MapReduce, users must transform and code them as customized map and reduce functions. One can identify the drawback of MapReduce is its lack of raised and declarative languages. Now a days, may of high-level languages are using MapReduce.

To, deliberate how to process query in the MapReduce framework. And then, this paper proposes the query process. It works on distributed and parallel file systems. To simplify the parallel processing, data is separated into equivalent portions. MapReduce splits the development of parallel programs falls into two stages: map and reduce. The map phase, each mapper loads a data portions from DPFS and transforms it into a list of key-value pairs. The key-value pairs are buffered as L local files, where L is the number of reducers. All key-value files are sorted by keys. In the reduce phase begin, When mappers finish their processing. The key-value files are shuffled to the reducers, where files from different mappers are combined together. For values with the same key, the user defined processing logic is applied by the reducer and a new key-value pair is generated as the result. Finally, the results are written back to DPFS.

V. SIMILARITY BASED TERNARY CLUSTER

The notion of ternary cluster based recognition query processing, using nearest neighbor classifier with the help of SVM and MapReduce programming model is proposed. Initially, the user takes input image pairs i.e., ternary cluster. The ternary cluster relations are fall into three nested clusters i.e., user, tag and data for each user are discussed in the above section-1. Each cluster produces a set of intermediate key value pairs (images) with the help of map functions written by users.

Now, first the mean value for each image is calculated i.e., intermediate key value with the help of image normalization process and feature extraction. Different authors are, Mathew Turk and Alex Pentland, 1991 [6] and Ting Shan, Brain, Lovell, C., and Shaokang Chen., 2006 [5] expanded the idea of face recognition. Each cluster consisting of intermediate key values associated with the same intermediate key and passes them to the Reduce function with the help of MapReduce library. In the same way remaining clusters are also having the same process.

Finally, the user of the reduce function accepts an intermediate key and a set of values for that key for each user of the data cluster. To compare the mean value with the target image of database with the help of image matching by calculating distance measure based on similarity .i.e., Euclidean distances with the help of Support Vector Machine.

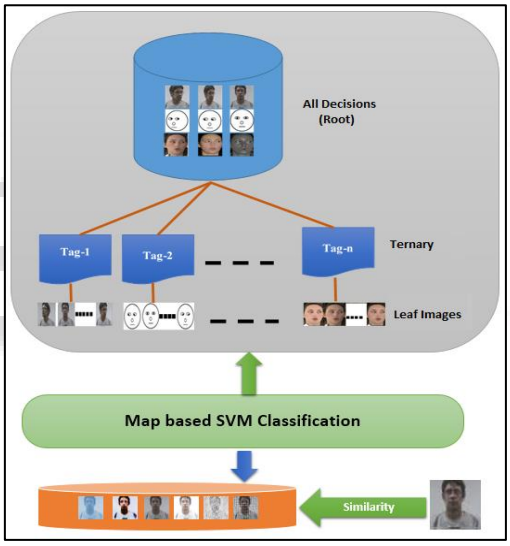

Fig. 2: Map Reduce with Support vector Machine based Foliage Image

VI. EXPERIMENTAL RESULTS

The experimental Results are produced on personnel signature cluster datasets by applying one of the approach analysis of variance which is to perform sample test whether the data belong to more than two or more clusters. The cluster software Reliability test is conducted on personnel signatures data sets to test for the equality of the means of two or more normal data set and also uses variances.

$$\textbf{Probability} = \frac{\textbf{Number of Failure Cases}}{\textbf{Total number of Cases}}$$

With support of this formula used to find software reliability by calculating the failure probability on the input datasets [9].

VII. CONCLUSION

Personally Identifiable system is used for De-Identification which provides the privacy for individuals. This system

incorporates large data sets i.e., Big Data by forming clusters. In hierarchical clustering, the output is a tree giving a sequence of clustering, with each cluster being a partition of the dataset. In automatic research of the clusters with the choice of the most relevant features are compared to the similarity process (Duplication). However, the system suffers from computational load on leaf images. To solve this problem this paper optimizes the automatic feature-subset selection based on T-nary cluster based support vector machine and also MapReduce technique is implemented. The main focus is to concentrate on foliage images i.e., particularly for a high-dimensional datasets and which are not at all relevant for a given operation. Also the low-dimensional feature subspaces are used to form the number of clusters and which are used to decide the cluster centers with the most relevant features at a faster pace.

REFERENCE

[1] Ajith Abraham, and Amit Konar and Swagatam Das, Automatic Clustering Using an Improved Differential Evolution Algorithm, Published in IEEE Transactions On Systems, Man and Cybernetics-Part A: Systems and Humans, Vol.38, No.1, January 2008.
[2] Isabelle Moulinier Feature Selection: A Useful Preprocessing Step. In Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research, P.P. 140-158, 1997.
[3] Jitao Sang, Changsheng Xu Senior Member, IEEE, and Dongyuan Lu, Learn to Personalized Image Search From the Photo Sharing Websites, IEEE Transactions on Multimedia, Vol.14, No.4, August 2012.
[4] Simon Tong, Daphne Koller Support Vector Machine Active Learning with Applications to Text Classification, Proceedings of ICML-00, 17th International Conference on Machine Learning, 2001.
[5] Ting Shan, Brain, Lovell, C., and Shaokang Chen., "Face Recognition to Head Pose from One Image" proceedings of the 18th International conference on Pattern Recognition, 2006.
[6] Turk, M., and Pentland, A. "Eigen faces for recognition", Journal of Cognitive Neuroscience, vol.3, no.1, pp.71-86, 1991.
[7] Vipin Kumar and G.Karypis, E.-H. Han. CHAMELEON: A hierarchical Clustering algorithm using dynamic modeling. COMPUTER, 32:68-75, 1999.
[8] Wiki-pedia.com/software reliability testing.

