Certificate Management System Using Fuzzy Based Clustering Approach

S.Benazir begam
Computer Science and Engineering
UCE-Thirukkuvalai
Nagapattinam,India
Benazirsyed1996@gmail.comm

M.Chandralekha
Computer Science and Engineering
UCE-Thirukkuvalai
Nagapattinam,India
chandralekha96@gmail.com

U.Preethi
Computer Science and Engineering
UCE-Thirukkuvalai
Nagapattinam,India
preethiudhaya.u@gmail.com

N.Radhika
Computer Science and Engineering
UCE-Thirukkuvalai
Nagapattinam,India
radhikavmd2013@gmail.com

Abstract— Big data storage is a computing model in which data is stored on remote servers accessed from the Internet. It is maintained, operated and managed by a cloud storage service provider on storage servers that are built on virtualization techniques. Big data storage can provide the benefits of greater accessibility and reliability; rapid deployment; strong protection for data backup, archival and disaster recovery purposes and lower overall storage costs as a result of not having to purchase, manage and maintain expensive hardware. There are many benefits to using cloud storage, however, big data storage does have the potential for security and compliance concerns that are not associated with traditional storage systems. We can implement the project that is centralized, fuzzy-based clustering certificate management platform that simplifies users to maintain all certificates with us. And certificates can be loss, if they keep in hand. So in this project we can design application for users. They can register into system and upload the certificates such as voter id,aadhar card, mark sheets and so on. Admin can extract certificate number and matched with database for predicting fraudulent activities. If it is occur means, the intimation is sent to crime department. User can view, download and print the certificates anywhere and anytime. This application can be user friendly and easy access GUI for all users.

Keywords-*big data; fuzzy based clustering;virtualization;*

*****

I.INTRODUCTION

Big data storage is a storage infrastructure that is designed specifically to store, manage and retrieve massive amounts of data, or big data. Big data storage enables the storage and sorting of big data in such a way that it can easily be accessed, used and processed by applications and services working on big data. In this module, we can create the system to handle big data storage. It has two controls such as admin and user control. Admin can be the responsibility to maintain all details about certificates. User can be register their details and get login for enter into the framework. Upload the certificates such as voter id, aadhar card, mark-sheets and so on.

A.Big Data

Big data is a term for data sets that are so large or complex that traditional data processing application softwares are inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy. Data mining involves finding interesting patterns from datasets. Big data involves large scale storage and processing (often at a datacenter scale) of large data sets. So, data mining done of big data (e.g, finding buying patterns from large purchase logs) is very interesting and is getting lot of attention currently.

B.ASP.NET

ASP.NET is an open-source server-side web application framework designed for web development to produce dynamic web pages. It was developed by Microsoft to allow programmers to build dynamic web sites, web applications and web services.ASP.NET's successor is ASP.NET Core. It is a re-implementation of ASP.NET as a modular web framework, together with other frameworks like Entity Framework. The new framework uses the new open-source

.NET Compiler Platform (codename "Roslyn") and is cross platform. ASP.NET MVC, ASP.NET Web API, and ASP.NET Web Pages (a platform using only Razor pages) have merged into a unified MVC 6.

II. RELATED WORK

The Big Data Analytics has been divided into three tiers. The first tier deals with the data accessing and computing second tier deals with the privacy considerations and the third tier deals with the data mining algorithms. The main problem in data mining is to generate global models by combining locally discovered patterns to form a unifying view. The Big Data is growing extremely, for high dimensional data the data reduction is important. The medical datasets are of high dimensionality in each field. The data reduction is easier for non-densed data rather than dense datasets. The large datasets can be done by combining clustering and classification. To obtain robust and stable clustering, consensus functions can be applied for clustering ensembles combining a multitude of independent initial clustering's. Direct applications of consensus functions to highly dimensional data sets remain computationally expensive and impracticable. Therefore a multistage scheme including various procedures for dimensionality reduction, consensus clustering of randomized samples, followed by the use of a fast supervised classification algorithm is needed. The ant colony optimization technique has emerged as a novel meta-heuristic belongs to the class of problem-solving strategies derived from natural (other categories include neural networks, simulated annealing, and evolutionary algorithms). The ant system optimization algorithms is basically a multi-agent system where low level interactions between single agents (i.e., artificial ants) result in a complex behavior of the whole ant colony.. A hybrid form of clustering which combines one or more clustering techniques can be used to cluster very large medical datasets. The four kinds of hybrid fuzzy cluster ensemble frameworks are used to cluster bio-molecular datasets. For preserving privacy in data-intensive applications, Twice-privacy algorithm based on utility matrix and multi-attribute clustering had been used. Twice –privacy conducts a clustering of sensitive values to protect similarity, sets different weight to retain quasi-identifier attribute to query service. In cloud environments distributed clustering which includes a novel distributed high dimensional data clustering algorithm based on Map-Reduce framework to distinguish the different communities from the entire social network had been suggested. K-Means algorithm had been proved to be better than FCM algorithm. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means because of the fuzzy measures calculations involvement in the algorithm. In fact, FCM clustering which constitute the oldest component of software computing are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction and it can provide approximate solutions faster. Single Pass (through the data) Fuzzy C-Means algorithm which is based on Weighted Fuzzy C-Means neither uses any complicated data structure nor any complicated data compression techniques, yet produces data partitions comparable to Fuzzy C-Means. Simple Single Pass Fuzzy C- Means clustering algorithm when compared to Fuzzy C-Means produces excellent speed-ups in clustering and thus can be used even if the data can be fully loaded in memory. Clustering technique on uncertain data (ie) clustering uncertain objects with the uncertainty regions defined by pdfs was difficult. For an accurate representation, at least thousands of sample points should be used to approximate an object's pdf. When applying the UK-means algorithm to cluster uncertain objects, a large number of expected distances have to be calculated. The basic min-max-dist pruning method is fairly effective in pruning expected distance computations. For finding the number of fuzzy clusters a new cluster validity index fwith crisp and fuzzy data had been suggested. The new index, called the ECAS-index, contains exponential compactness and separation measures. These measures indicate homogeneity within clusters and heterogeneity between clusters, respectively. Moreover, a fuzzy c-mean algorithm is used for fuzzy clustering with crisp data, and a fuzzy k-numbers clustering is used for clustering with fuzzy data.

III. DESCRIPTION & METHODOLOGIES

Big data are any data that you cannot load into your computer's primary memory. Clustering is a primary task in pattern recognition and data mining. We need algorithms that scale well with the data size. The former implementation, literal Fuzzy C-Means is linear or serialized. FCM algorithm attempts to partition a finite collection of n elements into collection of c fuzzy clusters. So, given a finite set of data, this algorithm returns a list of c cluster centers. However it doesn't scale well and slows down with increase in the size of data and is thus impractical and sometimes undesirable. In this project, we propose an extended version of fuzzy c-means clustering algorithm by means of random sampling technique to group the certificates. Certificates may be voter id, aadhar id, mark sheets or other important documents. Certificates are uploaded by users and implement automatic extraction approach to extract the certificate number from certificates. Admin can match the certificate number to database to predict fake user and restriction provide to upload the certificates. After that clustering can be done using Scalable Random Sampling with Iterative Optimization Fuzzy c-Means algorithm (SRSIO-FCM) to cluster the certificates

based on certificate number. Then implement security concept at the time of retrieving files from big data storage by using one time password. It is a password that is valid for only one at the time of download. OTPs avoid a number of shortcomings that are associated with traditional (static) password-based authentication; a number of implementations also incorporate two factor authentications by ensuring that the one-time password requires access to big data storage. And implement digital signature approach to embed the signature with date in certificate for additional reliability.

IV. RESULTS AND DISCUSSION

A.Fake Certificate Prediction

In this module we can implement the system to convert the text in image to readable format. After upload the certificates, we can automatically extract the number. This certificate number can be matched with database. If match is not found means, consider as fake certificate and reject to upload into cloud storage.

B.Certificate Clustering:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. In this module user can be upload original certificates. Certificates are grouped based on their certificate number and certificate name using fuzzy clustering algorithm. These clustering can be used for future purpose to retrieving easily. In this module, implemented a Scalable Random Sampling with Iterative Optimization Fuzzy c-Means named as SRSIO-FCM. It is a scalable model of RSIO-FCM with necessary modifications to tackle the challenges associated with fuzzy clustering of Big Data.

At the time of retrieval, user login to the system to view uploaded certificates. Search certificates based on their certificate number and certificate name. If the user need certificate means, admin can be One time password to user's registered mobile number. A **one-time password** (**OTP**) is a password that is valid for only one login session or transaction, on a computer system or other digital device. OTPs avoid a number of shortcomings that are associated with traditional (static) password-based authentication; a number of implementations also incorporate two factor authentication by ensuring that the one-time password requires access to something a person has as well as something a person knows (such as a PIN).After getting password, certificate should be download.

C.Embed The Digital Signature

This module provides intimation system to admin for download the files. Provide attestation using mouse pad to sign their signatures in canvas, this signature can be stored in database with date and time for future verification. Signature Pad is a jQuery plugin that takes advantage of HTML5 `canvas` element and java-script to create a flexible and smooth Signature Pad on your web page & app. Admin can capturing a signature during download increases security and decreases processing time for provide intimation to admin.

D.Certificate Retrieval

User can be download the certificate is in the form of black and white or color print. Attestation information can be stored in database. User can easily retrieve certificates anywhere and anytime. The framework of the proposed work as defined as follows:

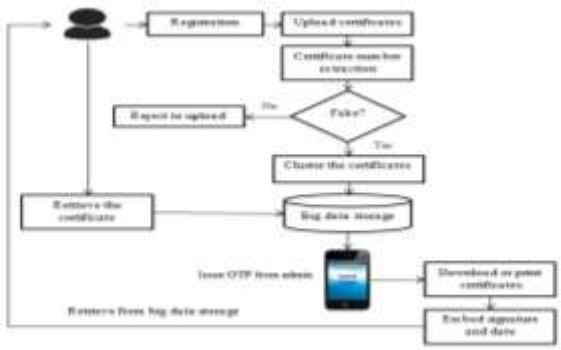

Figure1: system architecture

V.CONCLUSION

Big data refers to datasets that cannot be managed with current technologies or data mining software tools due to their large size and complexity. Big data mining is the capability of extracting useful information from these large datasets or streams of data. An effective way to cluster the large volume of data is Scalable Random Sampling with Iterative Optimization Fuzzy c-Means algorithm (SRSIO-FCM) by extending the clustering of sampled data to unstructured to structured datasets. The Distributed Environment has been set up where the very large datasets need to be reduced. The incremental fuzzy clustering can be enhanced with adding security using one time password and also provide embedded signature framework for improve the performance of big data storage.

REFERENCES:

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," pp. 1–137, 2011.

[2] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," Pattern recognition, vol. 41, no. 1, pp. 176–190, 2008.

[3] A. K. Jain, "Data clustering: 50 years beyond k-means," Pattern recognition letters, vol. 31, no. 8, pp. 651–666, 2010.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), vol. 31, no. 3, pp. 264–323, 1999.

[5] R. O. Duda, P. E. Hart et al., Pattern classification and scene analysis. Wiley New York, 1973, vol. 3.

[6] M. Steinbach, G. Karypis, V. Kumar et al., "A comparison of document clustering techniques," in KDD workshop on text mining, vol. 400, no. 1. Boston, 2000, pp. 525–526.

[7] Y. Zhang, S. Chen, Q. Wang, G Yu, "i2MapReduce: Incremental MapReduce for Mining Evolving Big Data," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 7, pp. 1906–1919, 2015.

[8] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-means algorithms for very large data," IEEE Transactions on Fuzzy Systems, vol. 20, no. 6, pp. 1130–1146, 2012.

[9] P. Hore, L. O. Hall, and D. B. Goldgof, "Single pass fuzzy c means," in Proc. IEEE International Conference on Fuzzy Systems (FUZZIEEE). 2007, pp. 1–7.

[10] P. Hore, L. O. Hall, D. B. Goldgof, Y. Gu, A. A. Maudsley, and A. Darkazanli, "A scalable framework for segmenting magnetic resonance images," Journal of signal processing systems, vol. 54, no.1-3, pp. 183–203, 2009.