Digital Document Streams Management System using Datamining Techniques

J. S. Balachander
Department of CSE
TRP Engineering College
Trichy

R. Ajay, T. Vigneshwaran, K. M.
Mohamed Hashim Aslam, J. Gowtham Raju
Department of CSE
TRP Engineering College
Trichy

***Abstract*— The principle purpose of using data mining is extracting data or information in large database. Document stream means continuous flow of document whether it may be pdf,word,images or any other types of document. we are proposing a method for retrieving top-k results for searching the digital certificates in cloud storage of an user. Generally we process the content in the certificate by extracting them using OCR(optical character recognition).Fuzzy clustering algorithm is implemented to cluster the certificates. And we provide security by using an OTP and digital signature is used for future reference. Fake certificate prediction is done based on certificate number. Scalable Random Sampling with Iterative Optimization Fuzzy c-Means algorithm (SRSIO-FCM) is used to cluster the certificates based on certificate content. User can search keywords in search box, based on that relevant certificates will be displayed. User can download the certificate after entering OTP number that is sent to the user's phone and email id.**

Keywords*— *Data Mining, OCR, Fuzzy Clustering Algorithm, SRSIO-FCM

I. INTRODUCTION

Today the use of big data is far behind the human knowledge and takes time to understand it. For instance a user is having a own storage with various types of files. It is hard for the user to retrieve the files when they needed it. By retrieving top-k results the user can search and retrieve it easily when required. So, we proposed this method for retrieving the results quickly.

In our model (DIGCERCH) , user must login into their id to save the information, records and other information, and upload the document. If user is new, then create new account for the user by entering the personal information. If the user uploads the certificate, it is verified with that certificate number. If It is original certificate then the certificate will be uploaded else it failed to upload. Certificate number is stored in admin database. Based on that fake certificate is predicted after checking it with admin database(govt database).

In existing model, clustering the big data is only possible with Fuzzy c Means(FCM). LFCM/AO - Implements to partition the data into clusters such as data-points in a cluster. Random Sampling plus Extension Fuzzy c-Means is used to train big data without delay. To classify the certificate, we used Scalable Random Sampling with Iterative Optimization Fuzzy c-Means algorithm (SRSIO-FCM) to cluster the certificates based on certificate content. It also predict the fake certificate and create alert system (i.e., successfully uploaded or failed to upload).

After uploaded in user DB, the certificate is clustered by using Fuzzy clustering algorithm. In DB, there is a large collection of documents. So, to find the document in DB takes more time (i.e., time complexity). For that we develop DIGCERCH. It searches through content based. User can search in the search box with top -k query documents, based on that relevant certificate is displayed. Time complexity is less when user search through DIGCERCH.

After finding the certificate, if the user wants to download user must enter OTP number which is sent to the user's phone and email id. It is for security purpose. The OTP is issued from admin. After entering OTP, relevant certificate is downloaded.

II. RELATED WORK

Top-k results are retrieved to the user over the web 2.0 by using a technique Partial order List (i.e.POL) . The principal purpose of this technique is to decreasing the cost of maintaining the sorted lists (i.e.arranged files) by grouping entries and ordering the entries only based on a fixed number of predefined boundaries instead of maintaining full order. These boundaries are then used to test the stopping condition. Continuous processing of data is possible by using this technique. We can retrieve the streaming of data from the web by using POL. But the major disadvantage is, large data collection. So at the time of retrieval , there may be a chance of getting irrelevant results. Due to large data collection,it takes more time to retrieve the relevant files (i.e. Time complexity) .

So to overcome the above drawback, threshold is used for efficient evaluation of Continuous Text Search queries. We can set the boundary condition to check the file. For eg: if we give 5 condition to check the file, it search and retrieves the particular files related to that condition. So limited document is only retrieved. The major advantage of this technique is, it provides ranked list of documents. Drawback of this method is, we can search and retrieve only plain text document.

Locally-optimal score bounds system is used for processing the continuous text queries. The process is done according to the similarity score given. It only process the homogeneous text. According to the given score, searching and retrieving of the file is done very accurately. Though it process large collection of data, the drawback is , it search only predefined document. We cannot upload the file what we want.

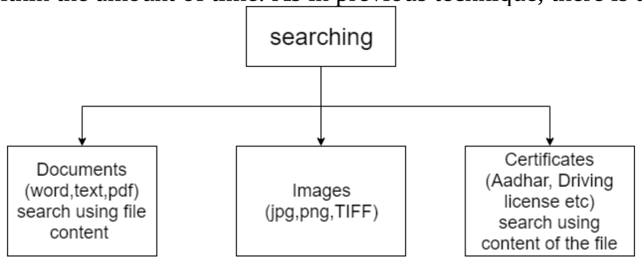
In general, ontology is a temporary DataBase. The result after searching the file is stored in temporary DataBase. It compartmentalizes the variable needed for some set of computations and establish the relationship between them. Various fields like artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science and information architecture all create ontologies to limit complexity and organize information. The ontology can then be applied to problem solving.we can search and retrieve the files easily from the temporary DataBase . Searching is done using the principle of auto complete pattern. Auto complete pattern is the feature in which an application predicts the rest of the word a user is typing. The file is searched according to the similarity of the keywords . Drawback of this method is, it can't be implemented in digital documents.

Log-Structured Inverted Indices (LSII) is work based on the following rule: it first search the relavant document according to the user query, then it stores the answer in the index and retrieves the document. For quick retrieval of the same document in the future, it store in the index. Microblogging allows the user to write brief text updates and publish them publically, through social media site or email.

II. PROPOSED MODEL

The above used techniques searches the file according to the file name only. It does not search the file content and it does not search the digital file. To overcome the above drawback, we proposed a system called DIGCERCH. It allows you to search the file content and also image content.

We implement this technique by using OCR.In previous technique, it searches only predefined document and retrieves the result . But our system allows you to upload any kind of files(i.e. Hetrogeneous Files). It will retrieves the relevant file within the amount of time. As in previous technique, there is a drawback, it takes more amount of time to retrieve the particular file. scalable Random Sampling with iterative optimization Fuzzy-c Means algorithm (SRSIO-FCM) is used to cluster the certificate based on certificate content.


searching
Documents (word,text,pdf) search using file content
Images (jpg,png,TIFF)
Certificates (Aadhar, Driving license etc) search using content of the file

Our system(i.e. DIGCERCH) predicts the user uploaded certificate is fake or not and we create a alert to indicate whether the certificate is fake or not. Fake certificate is predicted by using 2 steps of process. In first step, certificates are clustered based on certificate name and certificate number using Fuzzy clustering algorithm. In second step, Since all the certificate is stored in database(i.e.gove DB) it is checked with the DB. If the certificate number is matched then it shows the alert Record saved. If the certificate number does not match with the DB it shows the alert fake certificate.

Efficient keyword search is implemented to predict the relavant results. In addition to that, for protection to the user file, we provide OTP(i.e. One Time Password) to the user. After entering the OTP user will download the files which is sent to the user registered mobile number or user email id . In addition, we provide digital signature. The digital signature of the user is provide intimation system to admin for download the files.

A. OCR

OCR is nothing but Optical Character Recognition. It is a software that scans the digital documents and convert them into text format. When the page is scanned, it is stored as bit-mapped file in TIF format (i.e. Tag Image File Format). TIFF is a standard format for publishers and magazine layout.

OCR ALGORITHM

1. Loading any image format (bmp, jpg, png) from given source. Then convert the image to grayscale and binarize it using the threshold value.
2. Detecting image features like resolution and inversion.
3. Lines detection and removing.
4. Page layout analysis.
5. Detection of text lines and words. Here we also need to take care of different font sizes and small spaces between words.
6. Recognition of characters.
7. Saving the results to selected output format is necessary for instance, searchable PDF, DOC, RTF, TXT. It is important to save original page layout: columns, fonts, colors, pictures, background and so on.

REASON FOR USING OCR

1. To reduce Data Entry Errors
2. To Consolidate Data Entry
3. Human Readable
4. Scanning Corrections

B. MODI

We use MODI (i.e. Microsoft Object Document Imaging) to scan the digital text. It is a OCR application which enables editing and illustrating documents scanned by MODS (i.e. Microsoft Object Document Scanning). MODI can save text generated from OCR process into the original TIFF file.It checks in caseless comparison technique. However, digital document is scanned only when the image is clear in OCR. Hand written images cannot be recognized by OCR.

C. *DIGITAL SIGNATURE*

Digital signature is nothing but an attachment to any piece of electronic information, which represents the content of the document and the identity of the owner of that document uniquely. The purpose of using digital signature is because, it is used to validate the authenticate and integrity of a message, software or digital document. Digital signature is added to the certificates. The signature of the user is signed by the user using mouse pad in canvas. The signature can be stored in DataBase(i.e. DB) for future verification.

D. BENEFITS OF OUR MODEL

Our system (i.e. DIGCERCH) provides scalability for certificate management. There is a improved security for big data storage from the unauthorized user. Mainly it provides easily retrieval system to extract files from the system.

III. SYSTEM ARCHITECTURE

This system provides a cloud storage for the user to store their data. We provide a separate login id to each individual user. By this the user can upload their data. If they are not registered, they should provide their details in the registration form and then separate login id is generated for the user. The user's data may consists of any formats like txt,img,audios,video etc.

We create a database with certificate numbers (i.e govt database) for checking the fake certificates. When the user uploads a certificate, the certificate number is extracted using OCR. And it is checked with the database ,if the certificate number is matched then the file is uploaded.If the certifacte number doesn't matches, a alert message is given(i.e the file is rejected to upload).

After the certificate is stored in database, certificate is clustered using Fuzzy clustering algorithm. Scalable Random Sampling with Iterative Optimization Fuzzy c-Means algorithm is used to cluster the certificate based on certificate content.

It is hard for the user to retrieve the required data specifically from multiple files without knowing the file name.For this criteria, our system provides top-k results by searching the contents of the image. After matched with database, the relevant files will be displayed to the user.

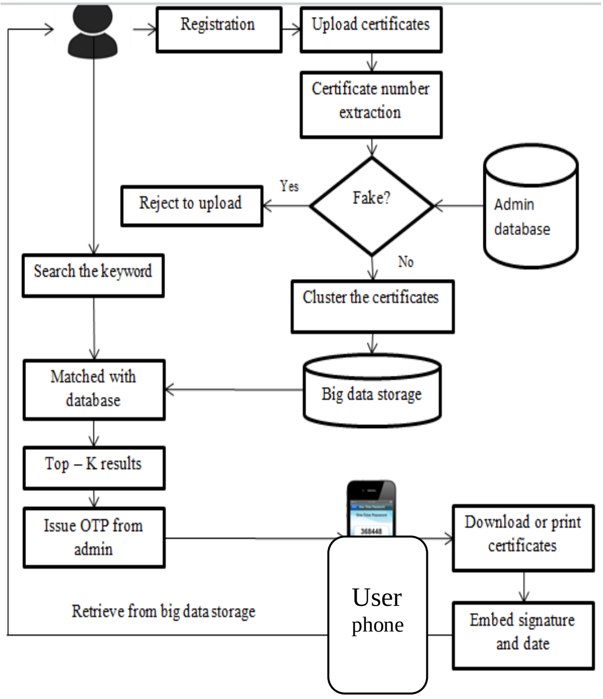


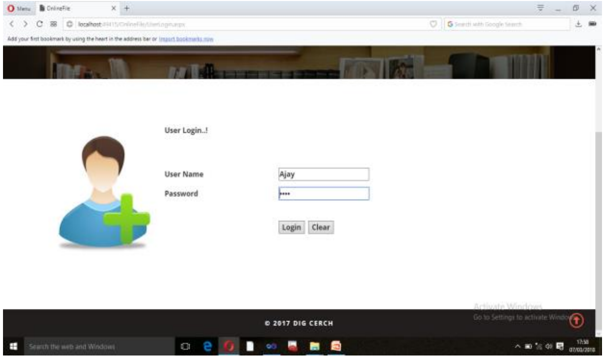
Figure: Digital Certificate Search

We provide security for the user to safeguard their files by providing OTP. If the user want to download their files, user must enter the OTP number which is sent to the user mobile and email. After entering the OTP, relevant certificate will be download.

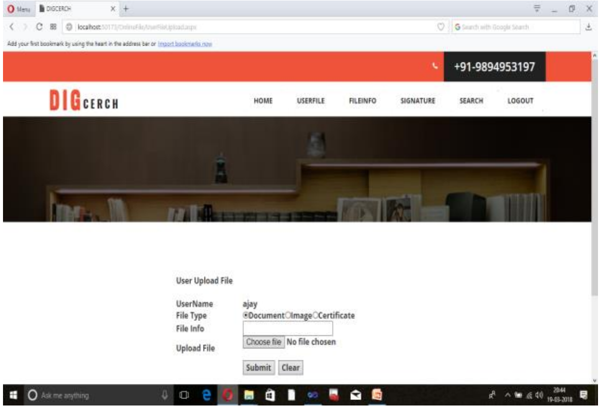
We provide attestation using mouse pad to sign user signatures in canvas. The signature can be stored in database(i.e. govt database) with date and time for future verification. Digital signature is a important application of public key cryptography. It serves the same purpose as of handwritten signature but is much secure.

IV. IMPLEMENTATION AND RESULT

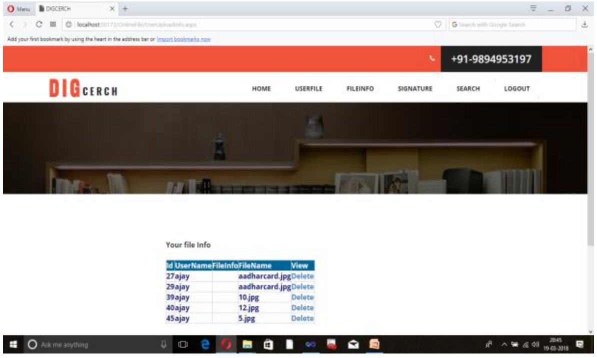
S.no	Input (user search)	Output (top k result)	Result
1.	Mohd	Aadharcard.jpg (certificate)	Success
2.	Aa	Aadharcard.jpg (image name)	Success
3.	Based	Abstract1.docx Template.pdf	Success
4.	1981	12.jpg	Success


User Login..!

User Name: Ajay
Password: ••••

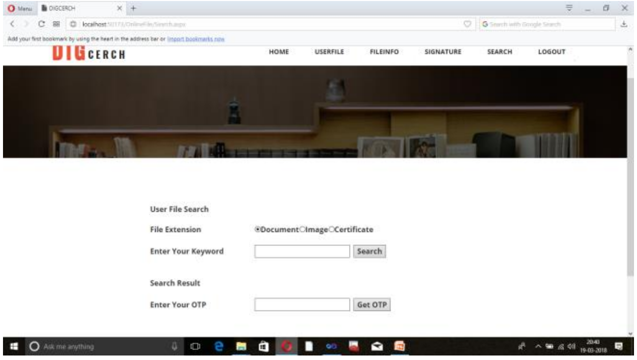

User Upload File

UserName: ajay
File Type: ● Document ○ Image ○ Certificate
File Info:
Upload File: Choose file No file chosen



Your file Info

Id	UserName	FileInfo	FileName	View
27	ajay		aadharcard.jpg	Delete
29	ajay		aadharcard.jpg	Delete
39	ajay		10.jpg	Delete
40	ajay		12.jpg	Delete
45	ajay		5.jpg	Delete



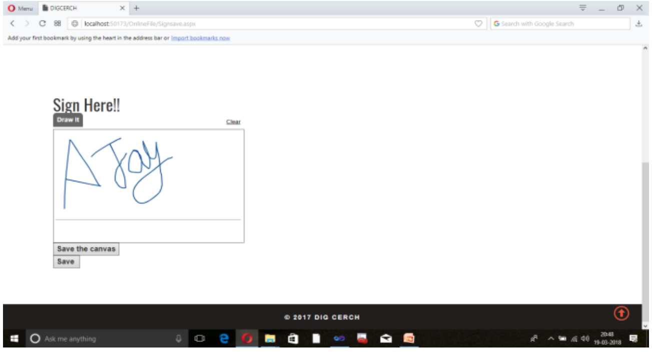
User File Search

File Extension ●Document○Image○Certificate

Enter Your Keyword

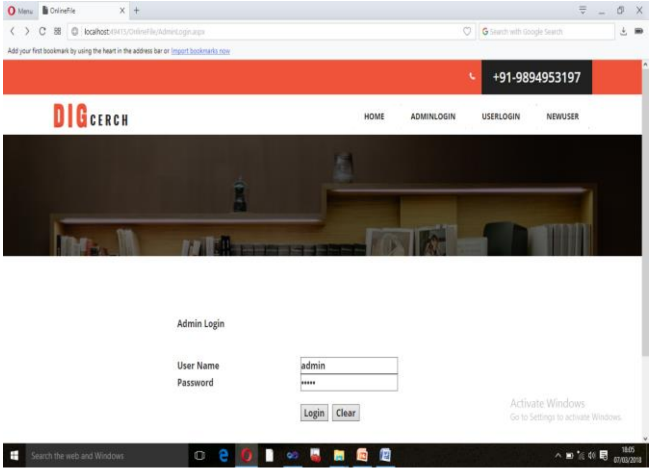
Search Result

Enter Your OTP



Sign Here!!

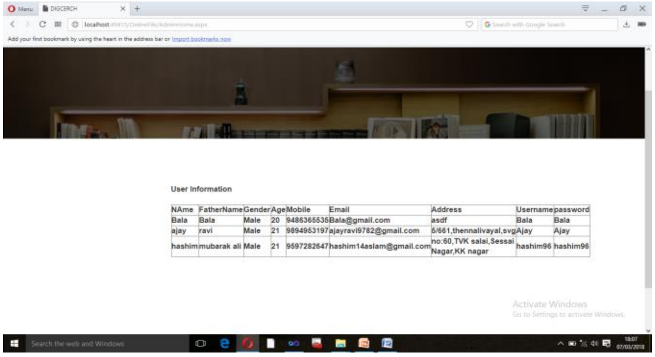
AJay



Admin Login

User Name admin

Password *****



User Information

NAme	FatherName	Gender	Age	Mobile	Email	Address	Username	password
Bala	Bala	Male	20	9486365535	Bala@gmail.com	asdf	Bala	Bala
ajay	ravi	Male	21	9894953197	ajayravi9782@gmail.com	5/661,thennalivayal,svg	Ajay	Ajay
hashim	mubarak ali	Male	21	9597282647	hashim14aslam@gmail.com	no:60,TVK salai,Sessai Nagar,KK nagar	hashim96	hashim96

VI. CONCLUSION AND FUTURE WORK

Our system is deployed in .net framework. The outcome of the process is 100% accuracy when the digital image is clear(i.e. HD IMAGES). DIGCERCH provides relavant file to the user and also provides security for the user file from the unauthorized person. In addition, we add digital signature of the user to provide intimation system to admin for download the files. From this model, it is clear that user can retrieve their files without knowing their file name. our future work is to process in blurred image.

REFERENCE

[1] Wu, L., Lin, W., Xiao, X., & Xu, Y. (2013).” LSII: An indexing structure for exact real-time search on microblogs”. *Proceedings - International Conference on Data Engineering*, (April 2016), 482–493.

[2] Hoppe, A., Nicolle, C., & Roxin, A. (2013). “Automatic ontology-based user profile learning from heterogeneous web resources in a big data context”. *Proceedings of the VLDB Endowment*, *6*(12), 1428–1433.

[3] Processing, P. T., Haghani, P., Michel, S., & Aberer, K. (2010). The Gist of Everything New : Categories and Subject Descriptors. *Cikm*, 489–498.

[4] Vouzoukidou, N., Amann, B., & Christophides, V. (2012).” Processing continuous text queries featuring non-homogeneous scoring functions”. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM ’12*, (February 2015), 1065.

[5] Mouratidis, K., & Pang, H. (2011).” Efficient Evaluation of Continuous Text Search Queries”. *Tkde*, *23*(10), 1469–1482.

[6] S. Prabhakar, Y. Xia, D. V. Kalashnikov, W. G. Aref, and S. E. Hambrusch, “Query indexing and velocity constrained indexing: Scalable techniques for continuous queries on moving objects,” IEEE Trans. Computers, vol. 51, no. 10, pp. 1124–1140, 2002.

[7] S. E. Robertson and D. A. Hull, “The TREC-9 Filtering Track Final Report,” in Text REtrieval Conference, 2000, pp. 25–40.

[8] Y. Zhang and J. Callan, “Maximum Likelihood Estimation for Filtering Thresholds,” in SIGIR, 2001, pp. 294–302.

[9] F. Fabret, H. Jacobsen, F. Llirbat, J. L. M. Pereira, K. A. Ross, and D. Shasha, “Filtering algorithms and implementation for very fast publish/subscribe,” in SIGMOD Conference, 2001, pp. 115–126.

[10] W. Rao, L. Chen, A. W.-C. Fu, H. Chen, and F. Zou, “On efficient content matching in distributed pub/sub systems.” in INFOCOM, 2009, pp. 756–764.

[11] M. Sadoghi and H.-A. Jacobsen, “Relevance matters: Capitalizing on less (top-k matching in publish/subscribe).” in ICDE, 2012, pp. 786–797.

[12] K. Pripuzic, I. P. Zarko, and K. Aberer, “Top-k/w publish/subscribe: finding k most relevant publications in sliding time window w.” in DEBS, 2008, pp. 127–138.

[13] A. Shraer, M. Gurevich, M. Fontoura, and V. Josifovski, “Top-k publish-subscribe for social annotation of news,” PVLDB, vol. 6, no. 6, pp. 385–396, 2013.

[14] I. F. Ilyas, G. Beskales, and M. A. Soliman, “A survey of topk query processing techniques in relational database systems,” ACM Comput. Surv., vol. 40, no. 4, 2008.

[15] M. Fontoura, V. Josifovski, J. Liu, S. Venkatesan, X. Zhu, and J. Y. Zien, “Evaluation strategies for top-k queries over memoryresident inverted indexes.” PVLDB, pp. 1213–1224, 2011.

[16] K. Mouratidis, S. Bakiras, and D. Papadias, “Continuous monitoring of top-k queries over sliding windows.” in SIGMOD Conference, 2006, pp. 635–646.

[17] A. Yu, P. K. Agarwal, and J. Yang, “Processing a large number of continuous preference top-k queries.” in SIGMOD Conference, 2012, pp. 397–408.

[18] W.Rao,L.Chen,S.Chen,andS.Tarkoma,“Evaluatingcontinuous top-k queries over document streams.” World Wide Web, pp. 59– 83, 2014.

[19] P. Haghani, S. Michel, and K. Aberer, “Efficient monitoring of personalized hot news over web 2.0 streams,” Computer Science R&D, vol. 27, no. 1, pp. 81–92, 2012.

[20] N. Koudas, B. C. Ooi, K. Tan, and R. Zhang, “Approximate NN queries on streams with guaranteed error/performance bounds,” in VLDB, 2004, pp. 804–815.

[21] R. Zhang, N. Koudas, B. C. Ooi, D. Srivastava, and P. Zhou, “Streaming multiple aggregations using phantoms,” VLDB J., vol. 19, no. 4, pp. 557–583, 2010.

[22] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” Commun. ACM, vol. 18, no. 11, pp. 613–620, 1975.

[23] G. Cormode, V. Shkapenyuk, D. Srivastava, and B. Xu, “Forward decay: A practical time decay model for streaming systems.” in ICDE, 2009, pp. 138–149.

[24] G. R. Hjaltason and H. Samet, “Distance browsing in spatial databases,” ACM Trans. Database Syst., vol. 24, no. 2, pp. 265–318, 1999.

[25] J. L. Bentley and D. Wood, “An optimal worst case algorithm for reporting intersections of rectangles,” IEEE Trans. Computers, vol. 29, no. 7, pp. 571–577, 1980.

[26] S. Ding, J. Attenberg, and T. Suel, “Scalable techniques for document identifier assignment ininverted indexes,” in WWW, 2010, pp. 311–320.

[27] H. Yan, S. Ding, and T. Suel, “Inverted index compression and query processing with optimized document ordering.” in WWW, 2009, pp. 401–410.