Criminal Identification System Based on Advanced Clustering Technique

Shahu Ronghe[1] Namrata Kandhari[2] Bhavin Jain[3] Sahil Shaikh[4] Trupti Dange[5]
[1]U.G Student [2]Assistant Professor
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]RMD Sinhgad School of Engineering

Abstract— Document clustering addresses the problem of identifying groups of similar documents without human supervision. In computer forensic analysis, thousands of files are examined. Much of the data in those files consist of unstructured text, whose analysis is difficult to performed. In such situation, automated methods of analysis are of nice interest. We present an approach that applies document clustering algorithms to forensic analysis that helps investigation. To automatically group the available data into an meaningful set of classes various clustering techniques will be used. It's seen that there is a huge amount of increase in the crime rate due to lack of technologies. Because of this there are many new opportunities for the development of new methodologies and techniques in the field of crime investigation using the methods based on data mining. Document clustering involves descriptor and descriptors extraction. Here, we are representing a model using new methodology for evaluation of document clustering of criminal database by using naïve bays and k-means clustering technique. This model clusters the criminal information basing on the sort crime which helps to police investigation and also we are generating graphs which are used to analysis of crime rate.

Key words: Sensors, Data Mining, Clustering, Information Retrieval

I. INTRODUCTION

In Document Clustering thousands of files are usually examined. Data in those files consists of unstructured text analyzing it by examiners is very difficult. Examining such complex and huge data is very difficult in day-to-day life. To handle such data many different techniques have been applied to solve the user query. Different use of Algorithm can be helpful in solving such query. By use of such techniques it can lead to easy and faster search of required data.

Today, collection and analysis of crime-related data are imperative to security agencies. The use of a coherent method to classify these data based on the rate and location of occurrence, detection of the hidden pattern among the committed crimes at different times, and prediction of their future relationship are the most important aspects that have to be addressed.

Clustering is a division of data into groups of similar objects. Each cluster consists of objects that are similar between themselves and dissimilar to objects of other groups. The main aim of document clustering scheme is to minimize intra cluster distances between documents, while maximizing inter-cluster distances. Clustering is form of Unsupervised learning and this is the only difference between clustering and classification. Clustering suggest groups based on patterns in data, whereas classification classify new sample into known classes.

The main purposes of crime analysis are:

- Extraction of crime patterns by crime analysis and based on available criminal information,
- Prediction of crimes based on spatial distribution of existing data and prediction of crime frequency using various data mining techniques, Crime recognition.

- In this paper, document clustering for criminal identification is implemented. For clustering of input document Naive-Bayes clustering technique is used. The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets.

II. LITERATURE REVIEW

The paper written by Ahmad AlShami, Weisi Guo, Ganna Pogrebna explores some of these challenges and tested the performance of two partitional gorithms for clustering such Big Urban Datasets. Two handy clustering algorithms the K-Means vs. the Fuzzy c-Mean (FCM) were put to the test. The purpose of clustering urban data is tocategorize it into homogeneous groups according to specic attributes. Clustering Big Urban Data in compact format represents the in-formation of the whole data and this can be net researchers to deal with this reorganized data much effciently.

The paper proposed by J Gao, ICFS tells us about an improved algorithm (ICFS) to deal with the several weaknesses of it. Unlike CFS, the proposed algorithm designs a formula for the cut off distance calculation and a method for cluster centers selection to improve its robustness.The ICFS method is evaluated on several datasets by comparison with the original CFS algorithm. Results demonstrate the effectiveness of the proposed method.

The paper written by KS Chaturbhuj gives the parallel clustering of large data set on Hadoop using data mining techniques. Hadoop can be used for processing such large data.

The paper proposed by Lubomir Stanchev can apply the k-means clustering algorithm to group the documents using a similarity metric that is based on key words matching and one that uses the similarity graph. We show that the second approach produces higher precision and recall, which means that this approach matches closer the results of the human study.

III. SUMMARY

A Dataset is created which will contain the data related to previous criminal record. All the records are clustered as per the crime category. Using this approach user will get easy

accessing of required data within few minutes. There would be easy search for new Police Inspector to get current information of cases in a particular area. Easy coordination is done between different branches of police station in a city. Also searching, retrieval, and Storage will be smooth and easy. Security of the document will be enhanced. The data can be represented in the statistical form such as bar graphs, pie charts, etc.

We have discussed several methods in the recent literature about document clustering. Proposed system is all about identifying the data using advanced clustering technique.

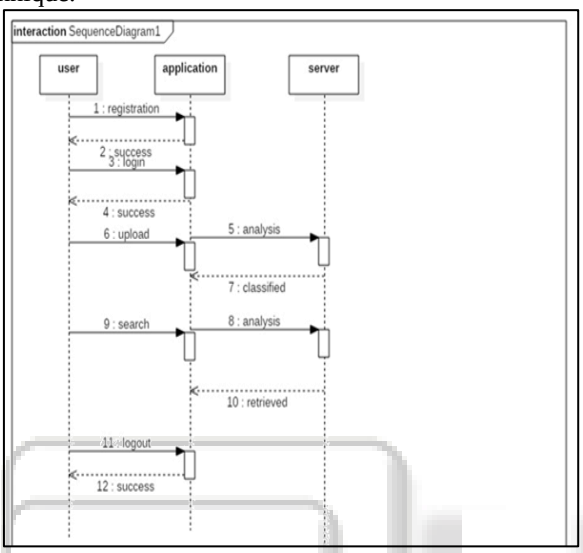

Fig. 1: System Sequence Diagram

The stages included in the proposed face detection system are:

1) Registration : Click on the registration button. Registration page should get open. Fill the form and submit.
2) Login: If password and username is correct it will successfully login.
3) Then fill up the new entry or can search for a past record on database.
4) Server finds out the information related to the person in the database with complete information and image and Displays it.
5) Finally required data is retrieved and logout.

The work flow of the system is described below

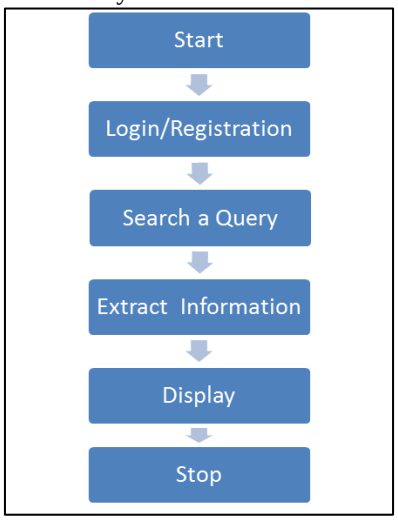

Fig. 2: Flow of the Experiment

IV. RESULTS AND PERFORMANCE ANALYSIS

Hardware Description: Intel Dual Core, 2GB RAM with Windows Operating system.
IDE: NetBeans.
Programming Language: JAVA
Database:SQL

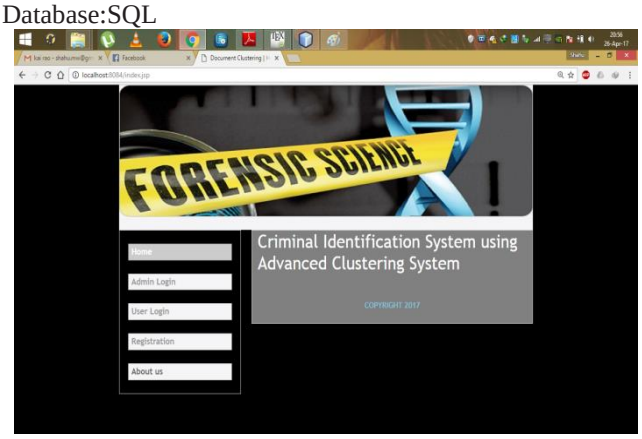

Fig. 3: Home Page

Fig. 4: Admin Login Page

Fig. 5: Criminal List

Fig. 6: User Registration

V. CONCLUSION

Collecting text document extract the information in that document in brief formats. It reduces the work of data examiner. It helps to police departments for storing the row data. This paper presents a new framework for clustering and predicting crimes based on real data. In this framework, different techniques like user image and different profiling identity to differentiate identity were used to reduce ambiguity of the data. The paper also proposed the easy and fast clustering of data.

The main purposes of the new framework for clustering and classification of crimes are mentioned below:

- Generation of training and testing data,
- Removing low-value attributes using weighting technique to deal with high-dimensional data challenge.

ACKNOWLEDGEMENTS

We would like to express our gratitude towards Mrs.Vina Lomte (HOD, Computer Engineering, RMDSSOE) and our guide Mrs.Trupti Dange for giving us an opportunity to review this paper.

REFERENCES

[1] MH. N. Gangavane; M. C. Nikose; "A novel approach for document clustering to criminal identification byusing ABK-means algorithm"2015 .
[2] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz, "The expanding digital universe: A forecast of worldwide information growth through 2010," Inf. Data, vol. 1, pp. 1–21, 2007.
[3] Sergio Decherchi, Simone Tacconi, and Judith Redi, Fabio Sangiacomo, Alessio Leoncini,and Rodolfo Zunino," Text Clustering for Digital Forensics Analysis", Journal of Information Assurance and security", Vol. 5,pp. 384-391,January 2010
[4] Dhanabhakyam, M and Punithavalli, M. A Survey on Data Mining Algorithm for Market Basket Analysis. Global Journal of Computer Science and Technology, Vol. 11 issue 11, version 1.0, 2011.
[5] Agrawal, R., Imielinski, T. and Swami, A. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data,207-216, 1993.
[6] L. Kaufman and P. Rousseeuw, Finding Groups in Gata: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.
[7] M. Agosti, F. Crestani, G. Gradenigo, and P. Mattiello. An Approach to Conceptual Modeling of IR Auxiliary Data. IEEE International Conference on Computer and Communications, 1990.].
[8] A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions," J. Mach. Learning Res., vol. 3, pp. 583–617, 2002.
[9] M. Agosti and F. Crestani. Automatic Authoring and Construction of Hypertext for Information Retrieval. ACM Multimedia Systems, 15(24), 1995.
[10] Semantic Document Clustering Using a Similarity Graph Lubomir Stanchev 2016 IEEE Tenth International Conference on Semantic Computing (ICSC) Year: 2016.

