ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol.2, Issue 12, December 2013

Clustering data using labels for disease prognostication and making meticulous firmness for fast recovery

E.Priya[1], C.Nivetha[2], A.Rajalakshmi[3]

Information Technology, Christ college of Engineering and Technology, Pondicherry[1,2,3]

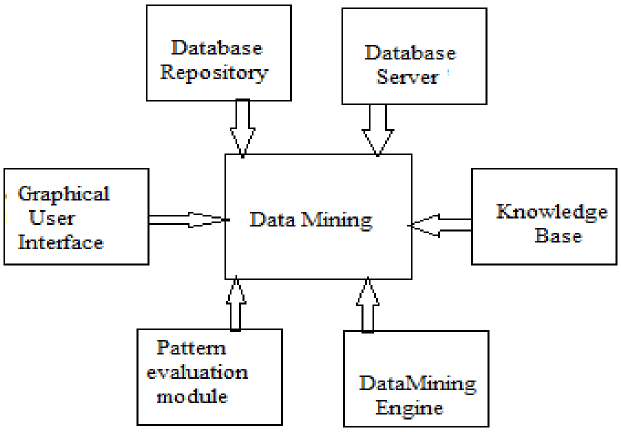
Abstract: Data mining is a process of searching through large amounts of data for patterns recognition. It is a relatively new concept which is directly related to computer science. It can be used with a number of older computer techniques such as pattern recognition and statistics. The goal of data mining is to extract important information from data that was not previously known. This paper make use of data mining concept for collecting patient details in hospital management .We use a new algorithm called RENOVATE algorithm to cluster the patient record for citation the disease of the patient. This algorithm cluster the data based on everyday update and produce the better result for the doctor to understand the state of affairs of the patient. The renovate algorithm provide the result by clustering the record on the core of labels. And also gives how the clustering process carried to place them in labeled order.

Keywords: clustering, data mining, Renovate, label, Prognostication.

I. INTRODUCTION

Data Mining is a process of extracting the frequent information or patterns from data in large databases. The actual data mining is the process of automatic or semi-automatic analysis of large set of data to retreive previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques[23].

There are many clustering algorithm available today to gather the data by comparing the similarities between the data and analyzing their distance matrix etc. The algorithm in use are hierarchical, agglomerative, ward's method and K-means etc. These algorithms perform clustering by calculating the outliers and similarities etc. But they have no advantage of clustering the large set of data with minimal time consuming and clustering in labeled order.



Block diagram of data mining

Database repository: A database of information about applications software which include author, data element, input, processor, output and interrelationships between the data. A repository is used as a CASE or application development system in order to find the object and business rules for reuse. It may also be designed to integrated third party CASE product.The repository is maintained by the certificate is valid, has expired or has been invoke

Database repositories are usually used and implemented in the data warehousing and business intelligence. This usually requires a level of aggregation of data that the lowel-level databases simply cannot provide, thus necessitating the creation of a higher-level structure[17]. A data repository is thus the logical aggregation of data items from separate databases into one centralized location for a specific purpose that cannot be achieved using the databases themselves[17].

Data base server: The term database server may refer to both hardware and software used to run a database, according to the context[17]. In case of software, a database server is the back-end portion of a database application, following the traditional client-server model. This back-end portion is sometimes called the instance[17]. It may also refer to the physical computer used to host the database.In the client-server computing model, there is a dedicated host to run and serve up the resources, typically one or more software applications[17]. There are also several clients who can connect to the server and use the resources offered and hosted by this server.

When considering databases in the client-server model, the database server may be the back-end of the database application, or it may be the hardware computer that hosts the instance. Sometimes, it may even refer to the combination of both hardware and software.

In smaller and mid-sized setups, the hardware database server will also typically host the server part of the software application that uses the database. In larger setups, the volume of transactions may be such that one computer will be unable to handle the load. In this case, the database software will reside on a dedicated computer, and the application on another. In this scenario, there is a dedicated database server, which is the combination of the hardware and software, and a separate dedicated application server.

A. Applications

Data mining is used in various applications includes business, fraud detection, banks, hospitals etc. In fraud detection it is used to find forensics related to the criminal case, in business it is used to find the customer activities and their satisfaction with the company products etc will help the company to improve their business process, in bank application data mining is used to cluster the customer records about their interest in particular policies.

B. Tools And Techniques

The Organizations that wish to use data mining tools can purchase mining programs designed for existing software and hardware platforms, which can be integrated into new products, or they can build their own custom mining solution. The data mining tool gathers the data, while the second program (e.g., the neural network) makes decisions based on the data collected[18].

(i) Traditional Data Mining Tools. Traditional data mining programs help companies to distribute the data patterns and trends by using a number of complex algorithms and techniques

(ii) Dashboards. Installed in computers to monitor information in a database, dashboards intimate the data changes and updates onscreen. This is displayed in the form of a chart or table

For instance, audit interrogation tools can be used to highlight fraud, data anomalies, and patterns.

II. CLUSTER ANALYSIS

It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Cluster analysis itself is not one particular algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their idea of what constitutes a cluster and how to efficiently find them. Popular ideas of clusters include grouped with minimum distances between the cluster members, dense areas of the data space, intervals or particular statistical distributions[22]. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results[22]. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to change preprocessing and parameters until the result achieves the required properties[22].

A. Clustering Algorithms

- ➢ Connectivity Based clustering
- ➢ Centroid Based clustering
- ➢ Distributed Based clustering
- ➢ Density Based clustering

The clustering algorithm are mainly based on the knowledge based work ,because in the database there are large data are present so to search the related data the knowledge based are important[17]. A knowledge base is a database used for knowledge sharing and maintenance. It promotes the collection, organization and retrieval of knowledge. A knowledge base is not merely a space for data storage, but can be an artificial intelligence tool for delivering intelligent decisions[17]. Various knowledge representation techniques, including frames and scripts, represent knowledge. The services offered are explanation, reasoning and intelligent decision support. Knowledge-based computer-aided systems engineering (KB-CASE) tools assist designers by providing suggestions and solutions, thereby helping to investigate the results of design decisions[17]. The knowledge base analysis and design allows users to frame knowledge bases, from which informative decisions are made. The two major types of knowledge bases are human readable and machine readable. Human readable knowledge bases enable people to access and use the knowledge[17]. They store help documents, manuals, troubleshooting information and frequently answered questions. They can be interactive and lead users to solutions to problems they have, but rely on the user providing information to guide the process. Machine readable knowledge bases store knowledge, but only in system readable forms[17]. Solutions are offered based upon automated deductive reasoning and are not so interactive as this relies on query systems that have software that can respond to the knowledge base to narrow down a solution[17]. This means that machine readable knowledge base information shared to other machines is usually linear and is limited in interactivity, unlike the human interaction which is query based[17].

Knowledge management (KM) contains a range of strategies used in an organization to create, represent, analyze, distribute and enable the adoption of experiences. It focuses on competitive advantages and the improved performance of organizations. Work script is a well known knowledge management database[17].

Data mining is the process of analyzing data from different perspectives and summarizing the data into useful information that can be used to increase revenue, cuts costs, or both[19]. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different ways or angles, differentiate it, and summarize the relationships identified[19]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases[19].

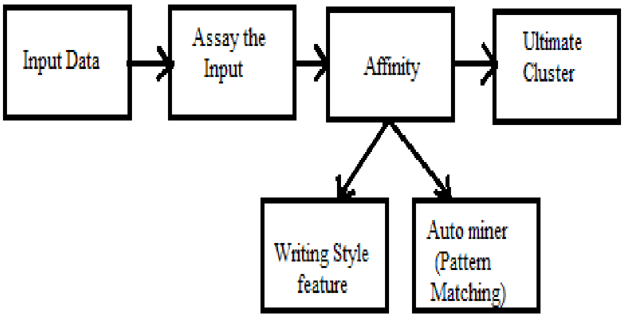
Frequent Patterns: as the name suggests patterns that occur frequently in data.

Association Analysis: from marketing perspective, determining which items are frequently purchased together within the same transaction.

III.EXISTING SYSTEM

In the existing work they use Computer forensic analysis. Data in seized computer consists of unstructured text, Algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis. This forensic analysis is done in order to find the evidence from the seized computer to provide it as a evidence in the court. Hierarchical clustering with various clustering algorithm is used to cluster the similar data that relates to case under analysis. When the data in the database are analyzed, they are clustered by calculating the minimum distance between the data and the similarities between them. The main cluster is identified by generating the dedrogram graph at time of distance calculation between the data.this can be done in case if any values are to be clustered.

If the data in the form of text are to be clustered in generating evidence for criminal case they will analyse and cluster the data by using tools called auto miner tool and writing feature styles tool. In these tools the writing style of the criminal are collected and the data in computer are compared with those styles being stored in the tool. If match occurs then those data are grouped into single cluster. The tool consist of writing style of e-mail by the criminal. Their usage of distance between the paragraph, their usage of sentence etc.



Block diagram of existing system

In this project we propose a detection of five diseases, their type and their suggested treatment. In hospital management hundreds and thousands of patient details are analyzed. Initially the data related to patients with different disease in hospital are collected and they are stored in the hospital database system. On the core of symptoms, test values and scan report the analysis is prepared . Each disease have some ranges for its test values to detect and symptoms to identify the type of disease. Based on the symptoms and ranges for the test report , the disease is identified and its type is detected by comparing their test values with the data present in the database. Once the type of disease is identified treatment is suggested for that particular disease. The clustering process is used in our previous work to group the similar data. In our proposed work we suggest the Clustering process with labels to prognosticate the disease by which the patient is exaggerated from and to produce them best suggested treatment.

IV.RELATED WORK

In the year 2005, Conan C. Albrecht Marriott School of Management Brigham Young University conan@warp.byu.edu the modern digital environment offers new opportunities for both perpetrators and investigators of fraud[20]. In many ways, it has changed the way fraud examiners conduct investigations, the methods internal auditors use to plan and complete work, and the approaches external auditors take to assess risk and perform audits. While some methods, such as online working papers, are merely computerized versions of traditional tasks, others, such as risk analysis based on neural networks, are revolutionizing the field[20]. Many auditors and researchers find themselves working amid an ever changing workplace, with computer based methods leading the charge[20].

Perhaps the most difficult aspect to computer based techniques is the application of a single term to a wide variety of methods like digital analysis, electronic evidence collection, data mining, and computer forensics. Indeed, **computer based fraud detection** involves a plethora of different technologies, methodologies, and goals[20]. Some techniques require a strong background in computer science or statistics, while others require understanding of data mining techniques and query languages. In the author's experience, discussions about computer based fraud detection techniques in most accounting circles are revolve around the use of **Ben ford's Law** to discover false invoices or other fraudulent amounts in corporate databases[20]. Analysis of data against Ben ford's distribution is useful, but it is only one of many computer based fraud detection techniques that should be used by professionals and researched by academics[20].

More **structured ways of working with these data are needed**. These methodologies can be included in textbooks and university courses[20]. There has been some work done on privacy issues, especially by the law field. However, more research is needed on **how to regulate data providers**, how to keep **information safe**, and where the lines of privacy should be drawn. As these web sites continue their move to international data, differences in country policies and cultures will become important research topics[20].

This paper reviews the different aspects of computer aided fraud detection. In particular, it describes each topical area, its research to date, and needed research for both professionals and academics.

In the year 2008, Farkhund Iqbal*, Hamad Binsalleeh, Benjamin C.M. Fung, Mourad Debbabi they propose a method and develop a tool for the investigator to visualize, browse, and explore the writing styles that are extracted from a collection of anonymous e-mails[25][12]. The relative strength of different clustering algorithms is evaluated. Our study reveals the relative discriminating power of four different categories of **stylometric features[25][12]**. Effects of the number of suspects as well as the number of messages per suspect on the clustering accuracy are addressed in this study.

The problem addressed in this paper is stated as: a forensic investigator has a collection of suspicious anonymous e-mails E. The e-mails are (presumably) written by K suspects, but the Investigator may or may **not know the number of suspects in advance[25][12].** The investigator wants to get an insight into the writing styles of an e-mail collection E, and **wants to identify major groups of writing styles called white prints** {WP1,.,WPk} in E.

Our objective is to develop a framework that allows the investigator to extract stylometric features from E and group e-mails E into clusters by stylometric features. In this paper, we propose a method and develop a tool for the investigator to visualize, browse, and explore the writing styles, found in a collection of anonymous e-mails E[25][12]. This study also focuses on evaluating different state-of-the-art clustering algorithms and determining which algorithms more suitable in a specific scenario.

Therefore, it is imperative to develop a **sound technique for keywords selection**. Features optimization will certainly be helpful in determining authors' style that is a true representative. Furthered, human behavior changes from context to context and from person to person. The need is to develop **methods for capturing style** variations for better authorship results[25][12]. **Addressing language multiplicity** is another research direction. The research of stylometric forensics is still in its infancy stage[25][12]. There is still a long way to develop a comprehensive, reliable authorship analysis approach before it can be widely accepted in courts of law.

In the year 2009, John Haggerty, University of Salford, UK ,Alexander J. Karran, Liverpool John Moores University, *UK* they proposed a tools such as EnCase and FTK to cluster the unstructured data[24].

Unstructured data is more complex as it comprises information about a social network, such as relationships, identification of key actors and power relations, and there are currently no standardised tools for its forensic analysis[24]. In addition, visualisation of this data can greatly aid the examiner in their understanding of the evidence.

Further work aims to develop the **framework and appropriate tools for digital investigations[24].** For example, there is a temporal element to unstructured data that would further enhance the understanding of social network information by identifying key nodes of influence and the development of the network over time[24]. In addition, work **aims to incorporate network narrative analysis tools**. In this way, the investigator may visualise not only the network relationships, but the structured data's influence on network behaviour[24].

In the year 2011, Sûreté du Québec, Montreal, Québec, Canada, they proposed a **Indirect relationship generation algorithm**.This algorithm is a hybrid version of both the open and closed discovery algorithmsIn this section, we present a method to discover the evidential trails between a prominent community identified in a dataset and other people in the document set who are not in the community. An evidential trail represents a relationship between the prominent community and other people through a common topic rather than co-occurrence. This trail is extracted as a chain of intermediate terms that link a community to a person

we have introduced the notion of prominent criminal communities and an efficient data mining method to **bridge the gap of extracting criminal networks**

information and unstructured textual data. Furthermore, our proposed methods can discover both direct and indirect relationships among the members in a criminal community.

In the year 2012, shravya lanka they proposed a data mining tools to cluster the required data in case of huge data sets[15].The tools they proposed are **Waikato Environment for Knowledge Analysis (WEKA)** are studied and the system uses WEKA to demonstrate the data mining methodology and thus retrieve data[15]. The four steps of data mining methodology including Association, Classification, Clustering and Regression are demonstrated on a set of data. Later, data retrieval is also performed using **Forensic tool Kit (FTK)** and the results are compared[15]. Retrieval of data is performed on storage device using data mining and compared to other forensic tools finally.

The rapid mining tool provides text mining, Text Mining is the process of deriving important data from large amounts of data. High quality data retrieval is called text mining. High quality data represents novel, interesting data which is relevant[15].

Some of the features of WEKA are it has forty nine data preprocessing tools, seventy six classification or regression algorithms, eight clustering algorithms, fifteen attribute/subset evaluator and ten search algorithms for feature selection.3 algorithms for finding association rules and three graphical user interfaces[15].

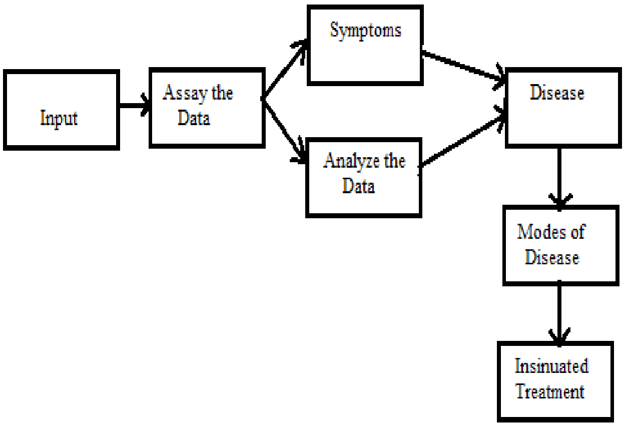
Therefore in this work the bank employee data set has been used to demonstrate WEKA data mining process in a step by step methodology while the data set containing data of Presidents of USA has been used for testing[15].

Thus it would be a future research area to **minimize cost and time** when smaller data sets are involved using data mining. Thus if this is implemented successfully, it reduces cost and time[15].

V.PROPOSED WORK

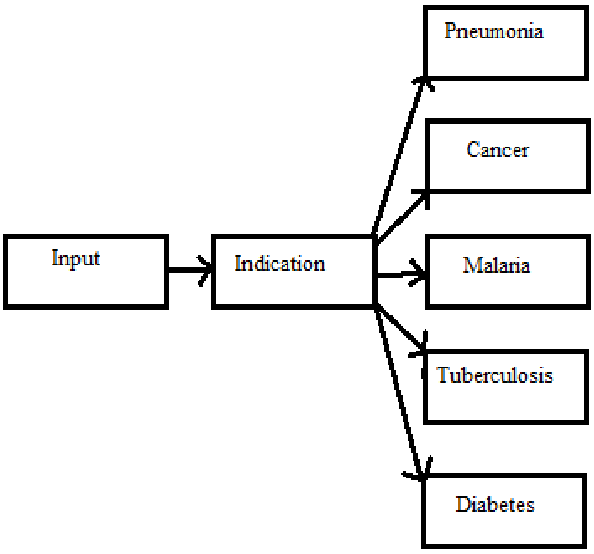
Details of patients in hospital is taken as an input and they are stored in the database of hospital management. These data are analyzed in order to predict the five disease namely (pneumonia, diabetes, cancer, tuberculosis, malaria) by which the patient is suffering from. Hence to detect the disease of the patient we compare the test results and symptoms of the patient and predict the disease of the patient. Based on the output it will provide the suggested treatment for the patients.

Labels are used in our work to identify the type of disease by which the patient is exaggerated from. For example if the patient have cancer, labeled clustering will help us to identify the type of cancer he/ she is suffering from. This detection is done on the basis of comparison between the test result ranges and the various symptoms of the patient. This clustering process is done for each and every update related to patient details. Clustering with RENOVATE algorithm is used to cluster the data on the basis of disease. Labels produce exact result about the disease, by which the patient is suffering from Result produced by labels give best treatment suggestion for the patients.



Block diagram of proposed system

REPORT ANALYSIS

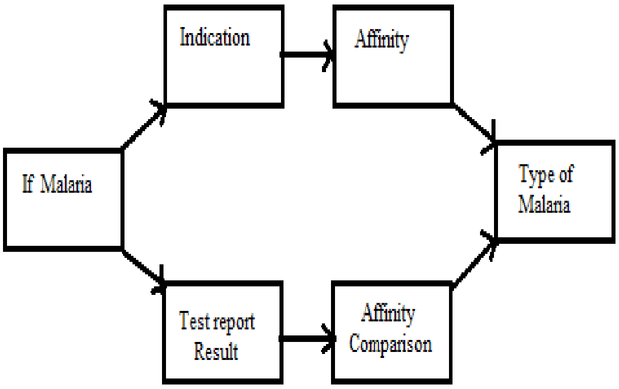


Block diagram of module 1

In this module we have to collect the data from the patients in the hospital in order to cluster the patient under five disease (pneumonia,cancer,typhoid,malaria,diabetes) the input data for these disease includes symptoms such as headache ,vomiting ,diarrhea,cause of jaundice,fast heart beat ,night sweat,dryness of skin ,weight loss ,blurry vision ,fever,cough and test taken for the disease includes chemotherapy ,CRT,MRI ,endoscopy ,blood test, ICT etc.

Based on the symptoms and test reports given for each disease cluster the patient data in the corresponding disease. The disease by which the patient is exaggerated from is produced as the result in this module.

DISEASE DETECTION

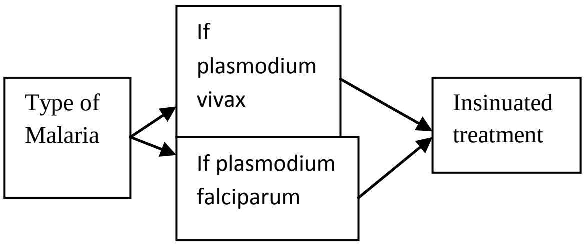


Block diagram of Module 2

Once the disease is identified based upon the symptoms and test result we have to predict the type of disease . For example, consider if the patient is suffering from malaria the immunochromatography test taken , if the blood level reaches PV line then we can say that he/she is suffering from plasmodium vivax or if the blood level reaches PF line we can say that he/she is suffering from plasmodium falciparum.

In the above example plasmodium vivax and plasmodium falciparum is the LABEL used here to cluster the patient details under the disease type of malaria.

DISEASE AND INSINUATED TREATMENT



Once the type of disease is identified we have to suggest the best treatment for the patient regarding the disease.

VI.ENVIRONMENT SETUP

This application implementation requires the installation of visual studio 2010(.NET), server management system. The database for the hospital management should be purged often when any patient record is no longer needed.

A.ALGORITHM

Consider the simple example to cluster the particular data from the set of data. Assume C,UT,BT as variables for cancer, urine test and blood test. If the UT and BT values is similar to range declared for cancer, then compare them to find which type of cancer they belongs to. Assume AAA as person with cancer. Compare the test report of patient to find the type of cancer from which he/she is suffering from.

```
For UT of i =1 do

{

C=the value of UT closest to cancer

If UT similariry(AAA,UT)<= threshold(10-20)

Then Say the patient is suffering from Benign;

Otherwise say patient suffering from Malignant;

}

If  BT similarity(AAA,BT)<=200-300

Person with benign

The value exceeds 300 the person suffers with malignant
```

```
if(@option = 'CheckDisease')
begin

select top 1 TestResultId
            ,[BloodTest]
            ,[UrinAnalysis]
            ,[BPTest]
            ,[SugarTest]
            ,[SputumTest]
            ,[HimoglobinTest]
            ,[Disease]
            ,[DiseaseType]
            ,[Instruction]
from TestResults
where [BloodTest] = @BloodTest
  and [UrinAnalysis] = @UrinAnalysis
  and [BPTest] = @BPTest
  and [SugarTest] =@SugarTest
  and [SputumTest] = @SputumTest
  and [HimoglobinTest] =@HimoglobinTest

end
```

This is a simple algorithm to calculate the similarity. In hospital management large set of data from database with thousands of disease with various symptoms are compared. This comparison is done to find the five disease and their type.

VII.IMPLEMENTATION AND RESULT

The database maintained in the hospital will consists of the input values depending on the arrival of patient everyday with various kinds of disease. If it is necessary to take the lab tests for the particular patient for the particular symptoms then the report of lab test is included in the database. And if the symptom of the patient belongs to the disease we are going to detect then it will be clustered under the type of disease on the core of given condition. Type of disease that we are in need to detect is said to be the LABEL. Therefore, the patient details will be clustered in the type of disease based on their test report and symptoms. Which will helps the doctor to give first inclination to the patient with crucial condition. The medication for the patient can be easily intimated by the doctor by examining the clustered report on the core of labels.

VIII.CONCLUSION

In this paper we have used the label based clustering, to find the exact disease of the patient and this clustering is done as soon as the update is made in the database it will provide us the current status of the patient and the treatment they are suppose to undergo. When the clustering process is done using labels it will produce the exact result which help us to make fast and correct decision about treatment to be given to the patients.

REFERENCES

[1] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*. London, U.K.: Arnold, 2001.

[2] 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.

[3] B. Mirkin, *Clustering for Data Mining: A Data Recovery Approach*. London, U.K.: Chapman & Hall, 2005.

[4] A. L. N. Fred and A. K. Jain, “Combining multiple clusterings using evidence accumulation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835–850, Jun. 2005.

[5] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, andM. S. Oliver, “Exploring forensic data with self-organizing maps,” in *Proc. IFIP Int. Conf. Digita lForensics*, 2005, pp. 113–123.

[6] L. Liu, J. Kang, J. Yu, and Z. Wang, “A comparative study on unsupervised feature selection methods for text clustering,” in *Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Engineering*, 2005, pp. 597–601.

[7] E. R. Hruschka, R. J. G. B. Campello, and L. N. de Castro, “Evolving clusters in gene-expression data,” *Inf. Sci.*, vol. 176, pp. 1898–1927,2006.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[9] 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.

[10] R. Xu and D. C.Wunsch, II, *Clustering*. Hoboken, NJ: Wiley/IEEE Press, 2009

[11] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D.Benredjem, “Towards an integrated e-mail forensic analysis framework,” *Digital Investigation, Elsevier*, vol. 5, no. 3–4, pp. 124–137,2009.

[12] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, “Mining writeprints from anonymous e-mails for forensic investigation,” *Digital Investigation, Elsevier*, vol. 7, no. 1–2, pp. 56–64, 2010.

[13] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, “Text clustering for digital forensics analysis,” *Computat. Intell. Security Inf. Syst.*, vol. 63, pp. 29–36, 2009.

[14] K. Stoffel, P. Cotofrei, and D. Han, “Fuzzy methods for forensic data analysis,” in *Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition*, 2010, pp. 23–28.

[15] shravya lanka “ Enhancing Forensic Investigation in Large Capacity Storage Devices using WEKA: A Data Mining Tool”,2012.

[16] Aggarwal, C. C. Charu, and C. X. Zhai, Eds., “Chapter 4: A Survey of Text Clustering Algorithms,” in *Mining Text Data*. NewYork: Springer, 2012.

[17] www.techopedia.com

[18] www.ukessays.com

[19] www.anderson.ucla.edu

[20] www.theifp.org

[21] www.studymode.com

[22] www.ijcsmr.org

[23] www.ijarcsee.org

[24] www.igi-global.com

[25] www.ncfta.ca