CLUSTER BASED ANALYSIS OF DIGITAL FORENSIC INVESTIGATION

By

Marriam Ghaffar

NUST201463925MSEECS60014F

Supervisor

Dr. Sharifullah Khan

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology

In

School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.
January 2018

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr/Ms __ Marriam Ghaffar __, (Registration No __ NUST201463925MSEECS60014F __), of School of Electrical Engineering and Computer Science (SEECS) has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members and foreign/local evaluators of the scholar have also been incorporated in the said thesis.

Signature: ____________________

Name of Supervisor: ____Dr. Sharifullah Khan____

Date: ____________________

Signature (HOD): ____________________

Date: ____________________

Signature (Dean/Principal): ____________________

Date: ____________________

CERTIFICATE OF ORIGINALITY

I here by declare that the research paper titled ***âĂIJCluster Based Analysis of Digital Forensic InvestigationâĂİ*** is my own work and to the best of my knowledge. It contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at SEECS or any other education institute, except where due acknowledgment, is made in the thesis. Any contribution made to the research by others, with whom I have worked at SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the projectâĂŹs design and conception or in style, presentation and linguistic is acknowledged. I also verified the originality of contents through plagiarism software.

Author Name: ___Marriam Ghaffar ___
Signature: __________________

DEDICATION

To All Those
Who focus on the light at the end of the tunnel, and Not on the length of the tunnel
To those
Who create stories that all of us aspire to have
But only a few have the courage, determination and will to create
Dedicated to three such souls

Abdul Ghaffar, Umair Rasul khokhar and Samina Ghaffar

ACKNOWLEDGEMENT

First and above all, I praise God, the almighty for providing me this opportunity and granting me the capability to proceed successfully. This thesis appears in its current form due to the assistance and guidance of several people. I would therefore like to offer my sincere thanks to all of them.

Dr. Sharifullah Khan, my esteemed supervisor, my cordial thanks for accepting me as a student, your warm encouragement, thoughtful guidance, critical comments, and correction of the thesis. He is the best supervisor I can ever have. I want to express my deep thanks to you for the trust, the insightful discussion, offering valuable advice, for your support during the whole period of the study, and especially for your patience and guidance during the writing process.

I warmly thank and offer my gratitude to Dr.Shahzad Saleem for his continuous support and excellent guidance. I also would like to thank Mr. Fahad Satti for his guidance in my research methodology.

I warmly thank and appreciate my parents for their material and spiritual support in all aspects of my life. I also would like to thank my husband for his assistance in numerous ways. I can just say thanks for everything and may Allah give you all the best in return.

ABSTRACT

Digital Forensic Investigation (DFI) is the investigation of crimes that involve the investigation of digital evidence, data and communication that are carried out on the suspectâĂŹs computers. DFI has become a research trend in the field of data mining because crimes ratio that carried out through computers is increasing. Moreover the essence of data mining in DFI is getting important because the capacity of computer storage and consequently size of data in computers are increasing with the passage of time. It becomes difficult to take the manual investigation of a computerâĂŹs data because it consumes too much time.In existing systems, data on crime scenes are retrieved from computers and clustered using clustering techniques on the basis of subjects defined by an investigator. The subjects are sensitive words related to the crimes. These clusters help in identifying relevant data on specific subjects which are useful for further investigation. The approach is also known as subject-based semantic document clustering.A drawback of the approach is that these generated clusters are concentrated on subjects, provided by the investigator and not on the subjects found from the suspectâĂŹs computer.

In this research we have also applied subject-based semantic clustering on documents found in the suspectâĂŹs computer. In order to resolve the above mentioned issue, the proposed approach first analyses documents and recommends subjects to the investigator for his selection. Then the investigator provides subjects for clustering of documents. The proposed approach applies overlapping clusters on the provided subjects and generates another generic cluster of documents that do not fall in the clusters of provided subjects. In addition, the generic cluster can be further passed to another cycle of this process for additional investigation. The experimental results show that the proposed approach provides comparatively more accuracy and flexibility than the existing systems.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Crimes have always been digitizing as the laws and laws enforcing agencies are getting advanced and strict. And the Forensic investigation required more effort and expertise. These are a lot of reasons behind it and in this chapter the digital forensic investigation and its needs are discussed. And it is explained why the devices and documents must be gone under forensic investigation and the motivation behind this topic is discussed. Now a days, in the world of fast computing all the things are being digitized. More things are being sold online now. The media has gone as social media from television or proper media. Crimes have also used the computing as a platform of committing crimes such as hacking, unauthorized access, child pornography etc. in this changing dynamics there is a need of some forensic tools for identifying and analyzing these crimes. That is why digital forensic is getting its popularity day by day in the field of crime investigation. But when the digital devices like cell phones, notebooks, laptops are found from the crime scenes then the digital investigation must go through the analyzing of documents stored in the suspects devices but as the storage is increasing and there are a lot of document stored in that storage. This activity consumes a lot of time for investigation. So using data mining technique for Digital Forensic Investigation has become popular area of research. Clustering is a technique of dividing data into the group of similar objects. Clustering is used in many fields, i.e. mathematics, machine learning, data mining and hence itâĂŹs a topic of research in different fields[1]. Clustering is one of the fundamental operation of Data Mining. Most appropriate and efficient algorithm for implementing clustering is k-means[2]. As K-means algorithm is effective and efficient when the data is large and its performance increases as the number of clusters increases [3]. .

1.1 Motivation and Scenario

There is a noticeable increase in the ratio crimes that are done by using computing technologies. The investigation of such crimes involves the investigation of digital evidences, data and the communication that is done by the suspectâĂŹs computer (Digital Forensic Investigation DFI). With the increasing storage of computer devices, the amount of data in a device have also been increased [4] and it takes time to investigate every data of the computer. Using data mining techniques, the DFI can be simple and less time consuming. Data from a suspectâĂŹs computer or any digital evidence that is found on crime scene can be grouped into the form of clusters using some subjects(keywords) as the basis of classifying the data into the clusters.

1.2 Problem Statement

The Present Subject Based Semantic Clustering Algorithm is an efficient model but it is still a time taking process. It creates Generic cluster which contains documents which are not covered under any previously defined subjects. Generic cluster might have documents that further provide investigator with insight into investigating crime and is then manually analyzed for this purpose. Manual browsing is a daunting and time consuming process for the investigator. The problem is how to minimize this effort of manual browsing and finding out the most relevant documents from the suspectâĂŹs device. Words that are most commonly used in those documents should be provided to the investigator.

1.3 Research Goals

The objective of this research is: âŮŔ To provide an accurate and flexible framework for DFI of documents found from criminalâĂŹs and suspectâĂŹs devices. âŮŔ To provide a framework that provides facility to the investigator in choose words, from the documents from suspectsâĂŹ device, that most frequently used. âŮŔ Moreover, the framework provide maximum data clustering by clustering the generic clusters

1.4 Proposed Methodology

The problem can be solved by using subject based semantic clustering on the documents found from the suspectâĂŹs device. The documents would be analyzed and then some suggestions would be offered to investigators. After the investigator selects some of the relevant terms then on those terms the documents would be grouped into the form of overlapping clusters of these terms as well as another cluster having other document which do not fall in any category i.e. Generic clusters. The algorithm will be designed for document clustering technique on which the documents are grouped into clusters. The model will start analyzing the suspectâĂŹs document through pre-process and generate the list of tokens from which it analyze what are the most frequent terms. Then these top frequent terms are given to the investigator as suggestions and the investigator can select the initial subjects. Then the subject formulation is done and the overlapping clusters of those subjects are created. In addition to those clusters, formed on a specific subject, a generic cluster, having all other documents, found on suspectâĂŹs computer, is created. Then the generic cluster is again go through all of the steps of subject-based semantic clustering to increase the accuracy of our approach.

1.5 Organization of Thesis

This section presents the overview of chapters and contents described in this thesis. This thesis is organized into six chapters. Chapter 1 is about Introduction of topic, problem statement, research goals and the motivation behind the study.Chapter 2 is on Background which briefly introduces the concept of digital Forensic, steps in digital Forensic investigation process, data mining and the main methods of clustering and classification. Chapter 3 presents a detailed state of the art of the area of the problem with its related information and related works with different aspects. Section 3.1 explains the role of data mining in crime investigation. Section 3.2 elaborates the different aspects forensic investigation using clustering techniques, through state of the art. Section 3.3 explains the Text and Documentation Clustering moreover it will explain the related work which explains this phenomenon. Section 3.4 is based on theories presented upon above section critically and summarizes them. Chapter 4 is based on the methodology of this research and the model that are made to answer the problem area. The whole process divided into 4 phases and the Section 4.1 to Section 4.4 presents each phase of Text and Documentation Clustering used for this research. Chapter 5 discusses the implementation. While chapter 6 discusses results drawn from the presented model and brief analysis of the results. Chapter 7 conclude the thesis by briefly discussing contribution, conclusion and future work.

2. BACKGROUND

For the past few years, computer forensics has grown to an increasingly important method of identifying and prosecuting computer criminals. Prior to the development of sound computer forensic procedures and techniques, many cases of computer crime were left unsolved. This is the reason Digital Forensic Investigation has gained such popularity within no time. Digital Forensic Investigation (DFI) is the process of investigating digital devices for the purpose of generating digital evidence related to an incident under investigation [5]. According to Carrier et al [6], digital evidence of an incident is any digital data that put some light about the incident.Digital evidence are favored in cases such as fraud, harassment, theft of trade secrets, or the more complicated cases such as homicides and many more where incriminating documents are likely to be found on the suspectâĂŹs computer.

2.1 Digital Forensic Research Workshop (DFRWS)

There are many digital forensic models that attempts to explain forensic process, one of them is Digital Forensic Research Workshop (DFRWS) [7] which is pioneer in describing the process. DFRWS describe this process as a linear process. The steps involved in the process are:

- Identification:

  It recognizes an incident of indicators and identifies its kind, suspected items, components and data associated with the incident. This element is very important because it affects other steps, but it is not clear in the field of forensic medicine.

- Preservation phase:

  In this phase, secure preserving digital evidences without damaging digital data being collected take place. It involves isolation, securing and preserving the state of physical and digital evidence.

- Collection phase:

  In this phase data is copied that might be related to the incident.

- Examination phase:

  This phase is aimed at facilitating the emergence of evidence while detailing the origin and significance. It involves detecting hidden and withholding information and related documents involving the conduct research in-depth methodology of evidence relating to the crime suspected [8].

- Analysis phase:

  Analysis involves determination of the significance, reconstructing fragments of data and drawing conclusions based on evidence found. This determines the significance and probative value to the case examined.

- Presentation phase:

  During presentation phase, summary and explanation of conclusions are presented.

2.1.1 Problems encounter using Forensic Tools

Examination phase is the most important phase of the investigation process. During recent years many researchers has put considerable amount of light on the issue of increasing amount of data in digital forensic field. Investigator has to go through humongous amount of data specially text documents to collect evidence and identify criminals. The process is sometimes dismaying due to large amount of documents found on storage. In keeping view of the problems investigator encounter due to large amount of data, investigator often uses forensic tools available in market to examine the collected data and perform an in depth systematic search for pertinent evidence. However there are some problems with these tools which are discussed below.

- High level Search

  Since surfing manually is a time consuming process and investigators often rely on the automatic search capability provided by either the DFI Tools menu or operating system to identify relevant evidence by searching documents found on suspectâĂŹs storage. Automated main research techniques provided by the DFIâĂŹs current research tools include regular expression search for approximate matching, search for keywords and the search last modified date. Unfortunately, this kind of techniques are applied directly against all document stored on computer without any prior knowledge about the topics discussed in each document. Therefore, the results based on these search techniques typically suffer from a large number of false positive and false negative results.

- Evidence-oriented Design

  Today's tools have been designed to help investigators to find concrete evidence, and not to help in the investigation. And assistance in dealing with offences committed against persons, when the documents located on your computer; they were not created to help resolve typical crimes committed with the computer or against your computer. Put roughly today's tools were created to deal with cases of drug dealing, not computer hacking cases [9]. They were created to find evidence, where possession of evidence in itself is a crime.

- Limited Level of Integration

  Many Forensic Tools today operate as a standalone application and have limited capability to work with other forensic tools already developed as a combined tool for investigation.

In keeping view of the problems DFI tools encounter, Data mining came into light in forensic field where investigator started using these mining techniques for solving crime cases. Data mining is a powerful tool which allows forensic researchers to search a large volume of data quickly and efficiently. It is used to extract information by discovering hidden relationships among data by using mining techniques out of usually large volume of data. Conventional data mining Techniques for example Association analysis, classification, prediction, and clustering and outlier analysis are used by law enforcement agencies in order to identify relationship in the data.

2.2 Classification

Classification techniques find common property between the different crime suspects and organize them into already defined classes. To predict crime trends, it demands predefined classifier for prediction and also a complete training and testing data because high missing values in data can affect prediction results. This technique is used to identify the source of e-mail spamming based on the structural features and patterns found in language used in email.

2.3 Clustering

Clustering techniques identify data items that are similar between themselves but different from the rest of the data. For example, it group suspects who commit crimes in similar ways and differentiate among criminals belonging to different criminal groups. Clustering is used to increase information retrieval during investigation.

Clustering technique is chosen over classification, since crime training data vary in nature and often contains several unsolved crimes. Classification technique that are based on existing and unsolved crimes, will not predict future crime with accuracy. Thus, clustering techniques work better to detect newer and unknown patterns from continuously increasing data.

3. LITERATURE REVIEW

Tools to commit crimes and other unethical thing have been spreading day by day which increased the need of Digital Investigation techniques to be more efficient [10]. This is why the literature on this area has wide range of researches. Some selected literature is described, in detail, in this section on the basis of sub-categories of this vast area.

3.1 Role of Data Mining in Crime Investigation

Now-a-days crimes are increasing widely due to which there is need of enhanced security. In police department the analyzed and clustered data of criminals was stored in criminal database. In [11] author Sukanaya .M determines the criminal tasks hotspot and find the criminals with the help of algorithms. Investigating and arresting guilty is the most problematic task because it requires comprehensive knowledge of crimes and criminals. Hotspot detection helps police to shun these activities in future. The classification is performed on the basis of crime type. Hotspot can be viewed by GIS. The application that based on intranet should be facilitated with the high security and not be accessible to unauthorized person. For protecting people effectively from crimes, criminal mapping is used which indicates where crime was occurred. Digital map helps to check the crime scene quickly. Crimes are classified on the basis of three attributes such as; 1) Classification on the basis of place, 2) Classification on the basis of crime type, 3) Classification on the basis of crime time. To place similar instances into sets the structured crime classification algorithm is used. For this purpose data clustering algorithm can also be used. The authors claimed that the techniques presented in this paper will help in the identification of hotspot which will led in decrement of crimes. But their introduced technique needs a lot of data about criminal activities and cannot be used for a certain scenario. We need an approach that help in for a certain criminal case too. Chen chung et al described the efficient and error free technique of mining crime data be useful for investigation by saving the investigator time so that he can invest[12]. They gave examples of different crimes such as traffic violation, sex crimes, theft, fraud cybercrime etc. they defined data mining as powerful and useful tool of investigation as it saves cost of time and hiring personnel. Some techniques of data mining are traditional like clustering, prediction, classification, sequential pattern mining or entity extraction. The authors found the relationship between data mining techniques that applied for crime investigation of above stated crimes in a graph. On the basis of these kind of literature and Tucson police department, they presented a general framework to analyze the crime data through data mining. Their framework determine the relation between different techniques of data mining that are applied on crimes (such as traffic violation, sex crimes, theft, fraud cybercrime etc.) and intelligence and criminal analysis. They focused on entity extraction, prediction, pattern visualization and association techniques. For association and prediction clustering can be effectively used while for pattern visualization and social networks analysis can be effectively done. They implemented their framework in coplink case study to demonstrate its application. They used the coplink data for research and used clustering and association technique to reveal the identities of cybercrime. Similarly, as in [11], this study helps in the classification of crime and categorizing of their types over data mining techniques. This research study could be extended by adding more dimensions about investigation not just of crime activities.

Richard Adderley inspected the contribution of data mining in crime scene investigator performance[13]. In this paper a new technique was presented to determine the performance of crime scene investigator which used computer based unsupervised learning algorithm. Data mining offers a wide range of techniques in various areas like data visualization, neural networks and statistic etc. cross industry standard process is a cyclic technique of data mining that was used in police north. The research was started by using insightful miner which is a data mining software. Nodes were placed on the worksheet that was developed on the needed process in order to achieve results. This research recommends CSI performance modeling was a practical selection for managers.The author describe the contribution and performance of data mining and described the process but did not discussed what is trending these days in this field and how can the existing processes can be improved. S. V Nath discussed the usage of K- means clustering algorithm that will help in detection and identification of patterns of crimes[14]. An attribute based weighting scheme was also developed that help others in enhancing law enforcement officers. After implementing K-means clustering algorithm the data was being prepared for investigation. The data is then transformed into de-normalized form by using extraction. Then checks are executed to check the quality of data. Detectives have a look at clusters and offers their recommendations. The limitation of this system is that it only helps detectives but canâĂŹt replace crime patterns and it does not provide appropriate information required for solving a certain case. The author Nicole Lang Beebe narrated that the digital forensic search aims to hunt each byte of evidence to trace desired string of text[15]. Due to large amount of data the investigator drowns into it and spend his time on irrelevant searches. For this problem there are two possible solutions available which are 1) Decrease the number of irrelevant search triumphs, 2) Present search hits in a manner which helps the investigator to determine related hits quickly. First approach was confusing for several investigator whereas the other approach was observed attractive. The present approach is based on the second solution. If some activities of text mining are applied on forensic text string searches then it will fit in the first solution. Those activities 1) Information Extraction: Detects conceptual associations by using syntactic outlines, 2) Content Summarization: shrinks the contents of documents, 3) Information Visualizations: graphically represents textual data, 4) Concept Linkage: recognize associations among documents. Their research seems to be complete and a knowledge contribution to the knowledge but it seems too complex and time consuming for implementation and usage.

3.2 Clustering Techniques for Forensic Investigation

Bide et. al. argued that, due to remarkable escalation in documents day by day, it is very necessary to isolate the documents of suspectsâĂŹ computers in appropriate clusters ([16]). Quicker categorization of documents is essential in forensic exploration but examination of these documents is very challenging. The authors presented an algorithm in which it is necessary to define number of clusters and its output is totally dependent on the provided input. The algorithm takes keywords as input and divides the document into small groups using divide and conquer technique. To resolve the issues of typical K-means algorithm, current algorithm was proposed. It uses divide and conquer and merge sort techniques on the documents. In the preprocessing uninformative words are removed. Preprocessing is performed before applying Vector Space Model which consist of 5 steps. These steps are Filtering, Tokenization, Stop word removal, Stemming and Pruning. Filtering eliminates punctuation marks from document. Tokenization divide sentence into words. Stop Word Removal eliminates words which have no meanings. Stemming shrink the words to its origin and Pruning eliminate the words with low frequency. VSM calculates the term frequency and number of words in a document. The input is distributed into small documents using divide and conquer to assemble parts of documents. The algorithm was tested on 20 new groups. A complete system was designed using all the models and the exactitude was measured according to F1 score. This algorithm takes less time than before. The conclusions of the experiment reveal that the F1 score is higher than existing algorithm and the time for clustering is also low. The approach of this study is appropriate for clustering and data mining of documents but the authors did not map their approach on digital forensic procedure.

Nassif et. al. proposed a method in which document clustering algorithm to forensic analysis was implemented[17]. The proposed method was explained by 6 famous clustering algorithms i.e. (K-means, Average Link, SingleLink, Complete Link, K-medoids, and CSPA) which were implemented to 5 databases. In the paper an algorithm was proposed which uses two indexes for validity which estimates about the total number of clusters. These experiment reveals that Average and Complete Link algorithm are most appropriate for this domain. Whenever any relating document is found, the forensic examiner performs the analysis on the documents that belongs to the area of interest. The algorithms mentioned above were executed by using the combination of its parameters. The execution of algorithm results in 16 vibrant algorithmic instances. The algorithm which were based on the SOM, clusters files on the basis of keyword search. Relevant application domain groups emails by using structural, lexical and semantics. Some preprocessing steps were taken before the execution of clustering algorithms. In present model every document is denoted by a vector which contains a frequency of word occurrences whose quantity was from 4 to 25. To determine the gap among documents they use cosine-based distance and Shtein distance. To predict the total number of clusters, an approach was used which takes partitions of dataset along with vibrant clusters after that. The required partitions were selected. The K-mediods algorithm is same as K-means. The only difference is that it computes mediods but not centroids. Hierarchal algorithms were executed and assess each portion from dendrograms by silhouette. If the selected. If the selected portion is singleton object then the process of clustering repeated recursively. The only limitation of this system can be its scalability. Silhouette was proved to be more precise than its modest versions. Our consequences recommends that using the file names with the document content information may be beneficial for cluster communal algorithms

Sergio Decherchi et. al. presented a methodology for text mining which was then applied via experiments[18]. The consequent investigation analysis was often done by time effort expensive human based analysis. The examiner did weighted examinations on contents gathered from forensic acquisition. In this task the text data established one of the central data cradles which may involve related information. Text extraction was the procedure which produces a pool of raw files that contains text. This procedure depends on powerful tool which handles huge data. The outcome of text mining helps in discovering patterns, tracking, classifying and extracting by using vibrant algorithm like detection and tracking etc. Extraction of textual data in majority relies on two techniques i.e. Meta data based and application based approach. In the first technique the detected files are recovered. Application based approach is when the metadata is smeared. Pieces of data are investigated with initials of header and footer. After that header and footer are investigated using same approach as of data. At the end the files that are not populated with text only were investigated, text is extracted and then that text was processed by using text mining tool. After performing text mining the documents were divided into clusters on the basis of similarity. D.B. Skillcorn proposed the ATHENS system design. ATHENS was general purpose system that elaborate huge information[19]. The ATHENS system finds terms from existing knowledge in massive data. ATHENS system starts with some terms that shows userâĂŹs current knowledge. By using the background knowledge of user novel contextualized queries are formed. The outcomes of these queries are clustered. After that the entire procedure is repeated. The major steps of ATHENS systems in this study are; 1) Closure (It specifies the content of keywordâĂŹs list. Here the set of nouns which are rated by their significance. It also enables the users to be casual.), 2) Probe (It performs the task of seeking of novel information from closure.), 3) Cluster (It organizes the resultant data of probe queries. The pages that donâĂŹt fit in any cluster are discarded.) 4) Iterate. The above mentioned phases are repeated many times. The validation of their system was problematic because the people who use ATHENS are ontological and the people suppose discrete page in a cluster to be similar only in a way that people would ponder similar. Their proposed system was designed to overcome both issues of piggybacking. The efficiency and effectiveness of this system are shown by in the corresponding results of their study.

3.3 Text and Documentation Clustering

Dagher and Fung introduced a subject-based semantic document clustering algorithm [10]. That will support the forensic investigation of the provincial police of SQ (Canada). In this digital forensic process the document and communication of a suspectâĂŹs computers are gathered in overlapping clusters on the basis of some subjects chosen by the investigator. The subject will be based on the area of interest of investigation, or which can identify any criminal activity. Their clustering solution used SVSM (Subject Vector Space Model). SVSM is used to represent documents, subjects or terms as vectors and demonstrate the relation between them. For building the definition of subject they proposed they proposed another scalable algorithm. The definition of subject will be based on initial definition in this algorithm. The initial definition is given as input from the investigator, each subject belongs to an area of crime investigation. This study can be very fruitful for the field of DFI but it can be simplified and extended. As it is not easy for investigator to estimate which subjects will be useful for clustering, there can be some other evidence providing documents that can be left due to selected subjects of investigator. It will be useful if the framework analyse documents and suggest some subjects for mining.

Anitha, G Thilagavathi also introduced a subject based semantic document clustering algorithm with the bisecting K-Means which can be used by investigator to group the documents into a cluster on the basis of a specific subject and a cluster of document which do not belong to any specific generic cluster [20]. They tried to increase the accuracy of this clustering technique has been improved by using K-Means. Bisecting K-Means is a combination of hierarchical clustering and K- means, used for the generic cluster. They also used SVSM for comparing terms with ESL Lookup, WordNet Synonym, and top frequent terms. They used WordNet to extract term synonym and word sense disambiguation to determine the sense of words. Then they used subject semantics clustering algorithm along with bisecting K means. The authors also used F- measure, precision and recall to improve the performance of clustering. The authors claims that they have made their presented technique unsupervised that no need for user to predetermine the subject and the clusters will made on the basis of subject. This research is quite complete and seems to be a good approach towards digital forensic investigation. But it can bring more accurate result if it will be used with F-measures.

Mascarnes et al also focused on the document clustering through which the investigator can cluster the document stored in a device found on crime scene systematically and it can provide subject suggestion for searching[21]. Their proposed DFI system can select subject both ways either by investigation or through suggestion given by system (via subject suggestion module). Subject-suggestion Module suggest the most frequent keywords for investigator to help his further investigation so that the investigator can select subject through suggested list or at his own. They implemented their framework using Java and NetBeans .According to the author's, subject suggestion can help investigator by selecting the appropriate query for searching hence it same time. Nicole L. Beebe used physical level text string search output relies on more than 2 million search triumphs that were found in about 50,000 allocated files[17]. Their study shows that LDA+ K-means uses centroid based user navigation process for best result. An experiment was conducted that used M57 patents datasets. The datasets involves daily images. Police seizure was specifically used. A query based on 36 terms was searched over police seizure. Four clustering algorithms LDA, SOM, K-means and kohonen for the experiment it is necessary to sustain the quantity of clusters constantly in order in order to enhance cross algorithm evaluation. Average information retrieval (IR) in both cases relevant and irrelevant triumphs were presented to users. The IR becomes an important deliberation in noisy search. User navigation was simulated by cluster navigation algorithm. Cluster navigation algorithm can be a part of 2 classes which are 1) Clusters are selected randomly but document are selected in each cluster with a sequence that starts from centroid. 2) Clusters and documents are selected randomly. They claimed that the above mentioned four algorithms enhances the precision rate of search hits up to 4.24

3.4 Critical Summary

Digital Forensic Investigation is a complex task and it is not limited to just one technique. Data mining is vastly used in DFI. It uses a number of different techniques for an effective and efficient investigation to reduce police work as described in [12]. The techniques of data mining are efficient for crime pattern detection[14][11][22], information finding for counter terrorism and intelligence [14][19], text clustering and string searching for finding important information hidden in the text documents[23][24][15][17][4][18] and forming the clusters on specific subjects for investigation [10][21][20]. The main focus of this study is about clustering of documents, found on a device of suspect, over a specific subject. This technique has been used in [10] but in this the clustering is done over the subjects given by investigator manually which is a daunting process. The problem of choosing subject by the suggestions is presented in [21] and it make a generic cluster of all other documents. But this generic cluster can have some sensitive terms, which can be useful for the further investigation. The study of literature has helped this research in refining the problem. And the direction of this research moves toward the subject-based semantic document clustering with the subject suggestion and re-grouping the generic cluster for a detailed investigation approach. âĂČ

4. PROPOSED METHODOLOGY

The problem described can be solved by using subject based semantic clustering as shown in figure 4.1 on the documents found from the suspectâĂŹs device. The documents would be analyzed after preprocessing and then some suggestions would be offered to investigators. After the investigator selects some of the relevant terms then on those terms the documents would be grouped into the form of overlapping clusters of these terms as well as another cluster having other document which do not fall in any category i.e. Generic clusters. The algorithm 1 presents the document clustering technique on which the generic documents are further grouped into clusters. The model will start analyzing the suspectâĂŹs document through pre-process and generate the list of terms from which it analyze what are the most frequent terms. Then these top frequent terms are given to the investigator as suggestions and the investigator can select the initial subjects. Then the subject formulation is done and the overlapping clusters of those subjects are created. In addition to those clusters, formed on a specific subject, a generic cluster, having all other documents, found on suspectâĂŹs computer, is created. Then the generic cluster is again go through all of the steps of subject-based semantic clustering to increase the accuracy of our approach. This chapter describes the overall description of the phases of the proposed solution in detail. For the realization of the system, Stanford CoreNlp, lucene, WordNet and given clustering algorithm is used that are java based libraries. NETBeans is used for developing application of the proposed solution as generic document clustering.

The algorithm of clustering documents is given below in figure 4.2 .

Phases of the proposed solution are as follows. Details of which are provided in subsequent sections.

Phase 1- Top Frequent Term: This is the first phase of proposed solution which computes repeatedly used terms from a document set after pre-processing, using the TF - IDF scheme. It involves the following steps:

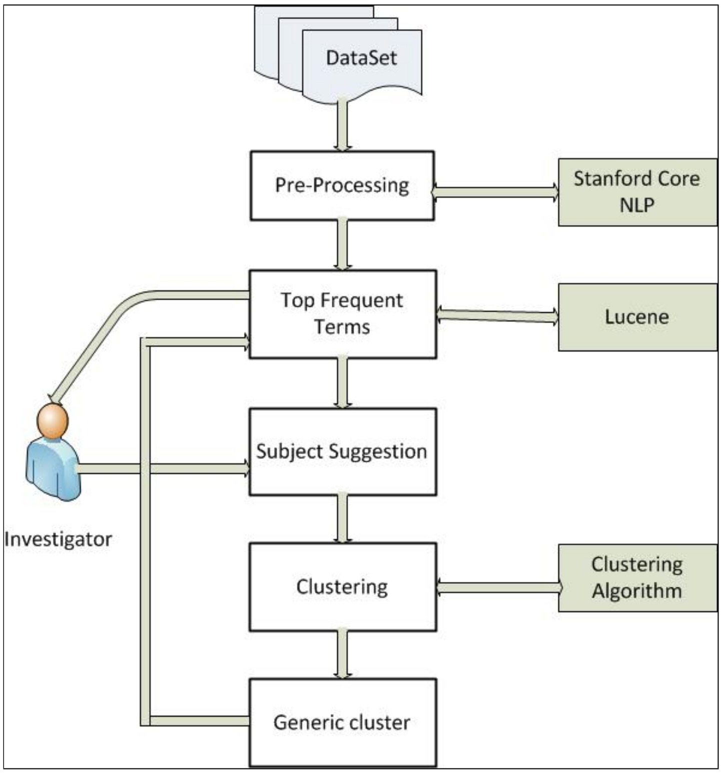
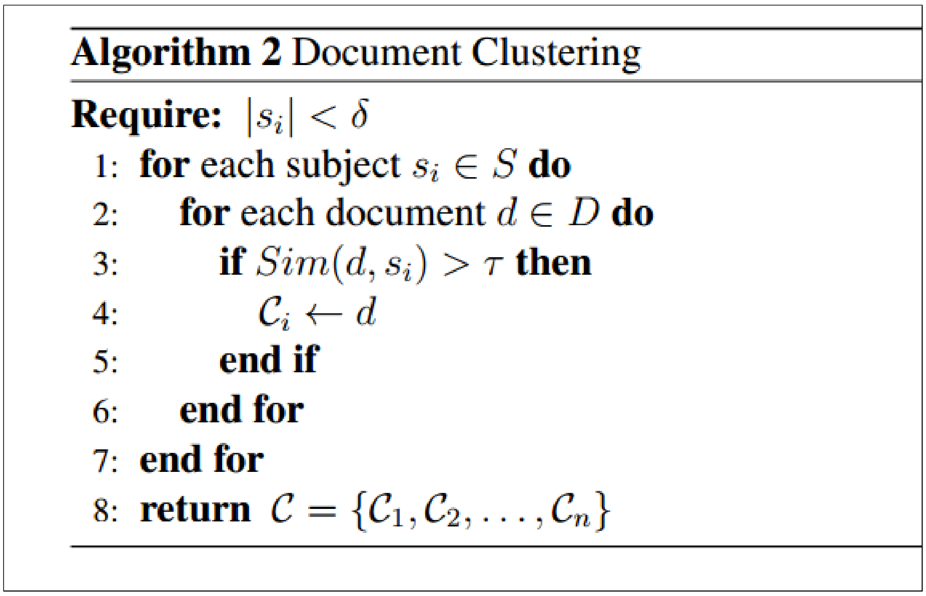


Fig. 4.1: An overview of the proposed solution



Algorithm 2 Document Clustering

```
Require: $|s_i| < \delta$
for each subject $s_i \in S$ do
    for each document $d \in D$ do
        if $Sim(d, s_i) > \tau$ then
            $\mathcal{C}_i \leftarrow d$
        end if
    end for
end for
return $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$
```

Fig. 4.2: An overview of the proposed solution

- Preprocessing
- Indexing
- Top Terms

Phase 2- Subject Vector formulation: Initial subject vectors are made by an investigator by selecting terms from a list of keywords. These subjects are further expanded by finding a synonym for each term in the subject from WordNet and selecting the most appropriate sense by using a Lesk algorithm for word sense disambiguation. This steps involves in this phase are following:

- Initial Subject Vector
- Subject Vector Expansion
  - Synset Repository Construction
  - Expansion using Synonyms
  - Synset Assingment

Phase 3- Document Clustering process: The clustering algorithm finds similarity between each document and every subject and based on this similarity generates an overlap cluster.

Phase 4- Re-clustering Generic Cluster: A generic cluster produced in the third phase is further divided into sub clusters by iterating through the previous phases.

Overview of the proposed solution is presented above as a flow diagram.

4.1 Phase 1-Top Frequent Terms

This is the first phase of the proposed model in which after documents are processed, top frequent terms are extracted from the entire list (Figure 4.3).

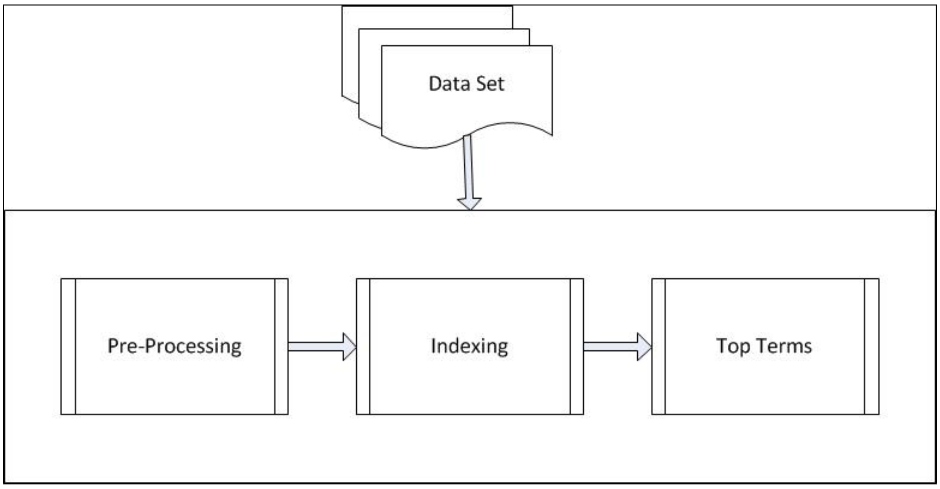


Fig. 4.3: 1st phase-Finding Top Terms

4.1.1 Preprocessing

This is the first step in which documents are processed to extract useful information by reducing noise and high dimensionality without significant loss in information. With this step, computational complexity to cluster documents is also reduced[1]. Three common text mining techniques namely tokenization, lemmatization and stop word removal have been used to preprocess the entire document set (Figure 4.4).

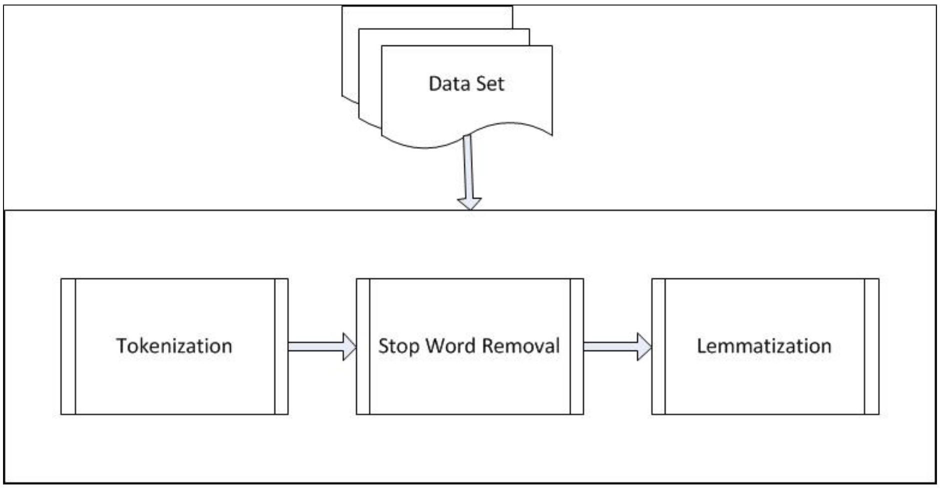


Fig. 4.4: Preprocessing from the 1st phase

Tokenization:

In lexical terms, tokenization is the process of breaking stream of text into words [25]. Phrases or characters produced are called tokens. Given a document set, tokenization uses whitespace, such as space or line break or punctuation marks to separate stream of text into tokens. A white space is not a part of the resulting token list. For example consider a character string: âĂIJit is rainingâĂİ, tokenization produces three tokens as âĂIJitâĂİ, âĂIJisâĂİ and âĂIJrainingâĂİ. Token list becomes an input for further preprocessing or text mining. Stanford Core Nlp [26] is used for the tokenization purpose. Document dataset is feed to the Stanford Core Nlp library which read each document in the data set in a sequence and in each documents it apply tokenization rules and produce tokens of the stream of text in a separate file of the same name.

Stop Word Removal:

Words that do not convey any information and are less important are removed from the list of tokens. For example preposition (he, she, it), conjunctions (and, or, but) which are considered common and do not play any role in providing information, is removed in this step. Stop word removal save space and increase search speed while indexing [25]. For instance consider a text string; âĂIJDigital investigation process is used to find evidence relevant to the incident under investigationâĂİ. After stop word removal, the resultant text string is âĂIJDigital, investigation, Process, evidence, relevant, incident, investigationâĂİ. Lucene stop word list is used to remove stop words while parsing a document set. The Lucene stopword list contains words that are common and does not convey any information. The parser reads the words in the document set and words that match with the word present in stop word list; parser removes them from the document set. Some common using Stop Words are in the table 5.14.

Tab. 4.1: Some common lucene stop words.

a	an	and	are	as
at	be	but	by	for
if	in	into	is	it
no	not	of	on	or

Lemmatization:

It is a process of reducing varied or morphological words to their base/root form generally a written word form so they can be analyzed as a single item. The process involves understanding context and determining the part of speech of a word in a sentence [25]. The reduced form of a word is known as lemma. For example Sleep, Sleeping, Slept is reduced to sleep after lemmatization. For example a text string found in a document is: âĂIJHeroin maintenance expands slowly in Europe: Heroin maintenance continues its slow spread in Europe.âĂİ After performing lemmatization output is âĂIJHeroin Maintenance Expand Slow in Europe: Heroin maintenance continues its slow spread in EuropeâĂİ. Slowly become slow, expands become expand and continues result into continue. Stanford Core NLP suit is used for the lemmatization process. Stanford Core Nlp while parsing documents read the tokens in the document set and lemmatizes them into their root form if needed.

4.1.2 Indexing

After Preprocessing unique terms from a document set create a list represented as T, are indexed and weights are computed for each term in every document. In indexing the list of documents against each term containing T is maintained for faster retrieval. After that, weight for each term is computed for every/each document. Hence, each document is represented as a vector of weighted terms. Weight is computed by using the term-frequency-inverse document frequency metric. TF is the frequency of the occurrence of a term in a document and IDF determines how much information the word provides. IDF is a measure of how important a term is. The Tf-IDF is used because it captures the global and local importance of a term when computing weights. For example, it not only considers the occurrence of a term within a document but also in the complete corpus. Specifically, each weighted term frequency is determined as

$$W_{t,d} = TF * IDF$$

Where

$$IDF = Log(1 + (|D|/Freq_{t,d}))$$

Hashmap is used for indexing terms. HashMap denoted as HashMap $< Key, Value >$ or HashMap $< K, V >$ maintains key and value pairs. HashMap implements the Map interface. Terms are saved as key in hashmap and weights computed are saved as the value. Hash Map enters the unique term only once but counts its occurrence that how many times a term repeats itself in a document. As words can occur many times in a document or document set, hashmap keeps the count information of a particular term in a specific document as well as a document set.

4.1.3 Top Terms

In the second step of the proposed model, top frequent terms that appeared repeatedly, from the entire list of unique terms are identified. In order to identify top frequent terms, the weight for each term is computed according to the formula above. The threshold value is set by an investigator as an upper limit to the number of top terms identified after this step. The term Document matrix (tdm) is created in which each column corresponds to a term t and each row belongs to a document d. Weights for each term in a document is calculated and the matrix entries are populated with these weights that are computed using hashmap in the previous step. With this matrix, where each entry correspond to weight of a term t in a particular document d, minimum threshold value suggested by an investigator, we then determine the top frequent terms whose weights are more than the investigator defined threshold value. Function for the top frequent terms can be defined as

$$Tft(M) = \{t \epsilon M | W_{t,d} > \delta\}$$

Where, M is a term-document matrix,$W_{t,d} = TF * IDF, \delta$ is a threshold value.

4.2 Phase 2: Subject Suggestion/Definition

In this second phase of the proposed system a complete extended subject vector is formulated which is the basis for clustering (4.5). This is the key phase of the model in which top frequent terms identified in the previous step are further analyzed by an investigator for subject formulation. By suggesting top frequent terms to the investigator, this phase aids the investigator to have insight into the suspect dataset and formalize the query accordingly. The purpose of the proposed model is to cluster suspectâĂŹs dataset according to the investigatorâĂŹs defined subjects denoted as S.

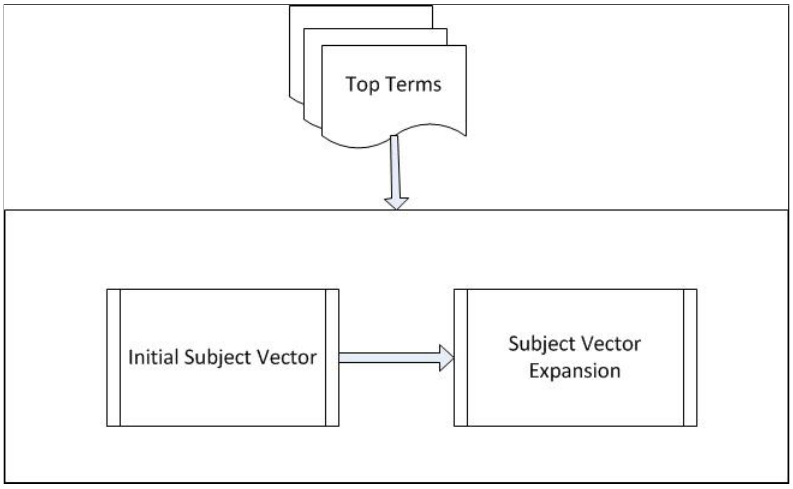


Fig. 4.5: Subject Formulation

4.2.1 Initial Subject vector:

For subject vector to made, an investigator is provided with a set of top terms. After examining top frequent terms, an investigator can select terms from the keywords suggested in the previous step that describe the subject vector denoted as $S_i = t_1, t_2$âĂę......t_n. An Investigator can make any number of subjects he is interested in.

4.2.2 Subject Vector Expansion:

Each subject is represented by a set of weighted terms that are related to the subject. These subjects are further expanded by finding synonyms from WordNet [27]. WordNet is used to find synset of each input term and Word Sense disambiguation is used to find the best synonym of the term. The input for this step is the set of weighted terms that an investigator select from the top frequent terms provided to him by the suggestion module. As each word has finite number of senses, finding synonyms for each input terms results in various senses. To find which sense is the best sense for the input term, word sense disambiguation algorithm is used. It is important to handle homonyms and polysemous words carefully by finding the best meaning of each term in the context. Java based Lesk algorithm has been used for this purpose. It finds the best sense of the terms by looking into the context of terms. Semantic expansion of the subject vector using WordNet involves three stages. In the first step, we define all the senses for each term. In the second step, we use Lesk algorithm to select the most suitable sense and finally in the last or third step we generate the expanded subject vector by allocating the synonym terms to subject vector terms (figure 4.6).

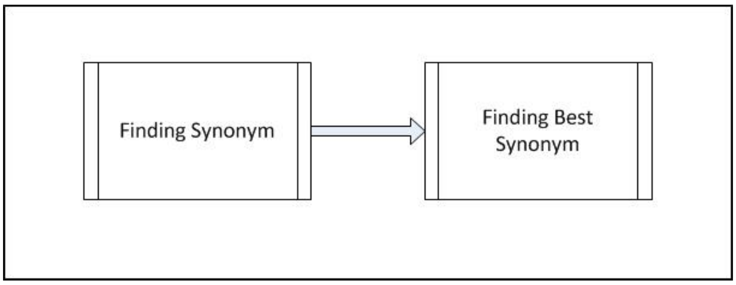


Fig. 4.6: Subject vector Expansion

Synset Repository Construction:

Let $w = t_1, t_2$âĂę......t_n be a set of input terms. Syn() defines a synonym function that takes a term t as an input and returns the synsets of both verb and noun from WordNet for the term t. Since WordNet lexicon contains majority of English words but if a special word occurs in the term list for example slang word then WordNet does not identify with the term and does not assign it with synonym.

Best Synset Assignment:

We get many terms T from a document set D but for each term it is not clear in which sense this term is being used in the document. Such a sense is called dominant sense and here we are going to find the dominant sense for each term t belongs to T. Each term extracted exists in multiple contexts say X. The main purpose of this step is to identify the best suitable synonym for each term t belongs to tf in the context of a sentence the term used in. To achieve this adapted version of LeskâĂŹs algorithm for word sense disambiguation has been used. This algorithm has been applied to each term to remove the following ambiguities:

1. Each context $x \epsilon X$ of the target term is defined as a frame of e term which appears to the right or left of the term in every occurrence.

2. For each context term, list all the senses searched from WordNet. item For the target term and for each context term list the following:

   (a) Its own WordNet gloss

   (b) The concatenated glosses of all all the synsets for all the context terms.

   (c) The terms are not tagged, so we associate with every term the synsets of its verb and noun senses.

   (d) Measure the similarity between each gloss of a target term with each gloss from each context terms by searching for overlaps.

   (e) Once each combination has been recorded, allot the synset of the matching sense with the highest score to the target term.

   (f) Repeat this process for every term until the most appropriated sense has been selected for each input term.

Finally after this, we find the dominant one among all senses.

Expansion Using Synonyms:

After allocating all the synonym set S to each term t, the subject expansion vector can now be created by giving the best synset of each term t belongs to w. WSd(t) gives the best synset for the input term.

4.2.3 SUBJECT VECTOR SPACE MODEL (SVSM)

SVSM is used to represent a subject vector S and a document set D in the space [10]. The subject vector space model is an algebraic model which is based on Vector Space Model (VSM) and the Topic Based Vector Space Model (TVSM). It is an n-dimensional model. In this model, each dimension or axis represent a subject such as si belongs to S. All axis coordinates in SVSM are positive just like in TVSM and also all axes are orthogonal to each other.

4.2.4 Document Subject Similarity Function:

Document subject similarity function defines the alikeness between a document and a subject and returns the positive value residing between 0 and 1. The coordinate value of a document vector in subject dimension is the similarity between subject and the document. Cosine similarity function is used as a similarity metric between subject and document. Formula for cosine similarity is observed as

$$Cos\theta = \frac{A.B}{|A|.|B|}$$

4.3 Phase 3: Semantic Clustering Algorithm:

This is an important phase of the proposed model. At this point of time, documents are processed and final expanded subject vectors have been formulated. Semantically clustering a document set D based on investigator defined subjects S is done in this phase. The clustering algorithm involves two stages. In the first phase, it generates a set of overlapping cluster after measuring the similarity between each document di belongs to D and each subject si belongs to S. A generic cluster is also created during the process having documents not related to subjects on which clustering done. In the second phase Generic cluster is further clustered using the same methodology to extract useful keywords and clustered according to subjects suggested by investigator. Clustering Algorithm used for clustering documents is referenced from [10] and is described above. For the entire document set, similarity is calculated between each document and subject, and the document is assigned to the cluster when similarity is above than the threshold i.e. δ. Flow chart for the algorithm used is shown in figure 4.7.

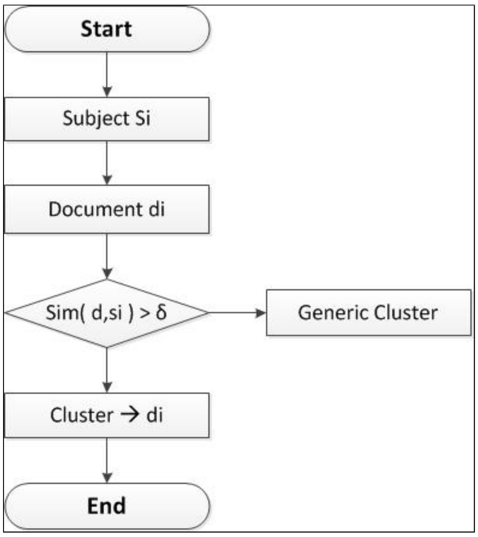


Fig. 4.7: Document Clustering Algorithm

4.4 Phase 4:Re-clustering Generic Cluster:

This is the last step of the proposed solution, in which the generic cluster generated as the result of document clustering in phase three is re-clustered through the previous steps of finding top frequent terms, subject vector formulation and semantically clustering documents. The input to these phases is a generic cluster instead of whole document set. A generic cluster contains many documents and manually browsing them is time taken. As it does not belong to any of the already given subjects an investigator does not know what important information resides in this generic document cluster. Hence it is important to find subjects to which documents in a generic cluster belong to. For this purpose, generic cluster is further processed by finding top frequent terms, subject formulation and clustered according to newly formed subjects. Top frequents terms are computed for the documents in a generic cluster; these terms are further expanded by WordNet and disambiguates through Lesk algorithm. After expended subject vector is created, generic cluster is semantically clustered using subjects formed in this phase. By re-clustering this generic cluster of documents, investigator came to know about the kind of data and hidden information lie in these documents of generic cluster.

5. IMPLEMENTATION AND EVALUATION

This Chapter will discuss specification for system and software as well as the Output of the proposed system, for illustration screenshots of the system are provided The system specifications, for the proposed system, are shown in table 5.1.

Tab. 5.1: Accuracy measures of each step.

Processor	Intel 1.7GHz Corei3 4010U 32-bit or 64-bit
RAM	4 GB
Operating System	Windows 8.1 Pro
Hard disk space	16 GB

The specifications of software, used in the development of the system, are shown in table 5.2. We used NetBeans IDE (Integrated Development Environment) to develop our proposed system as it is open source and has many built-in libraries useful for our proposed system. As discussed in chapter 4, the proposed system involves four phases, in which first phase includes preprocessing and extracting top terms from the processed documents (Figure 4.1).

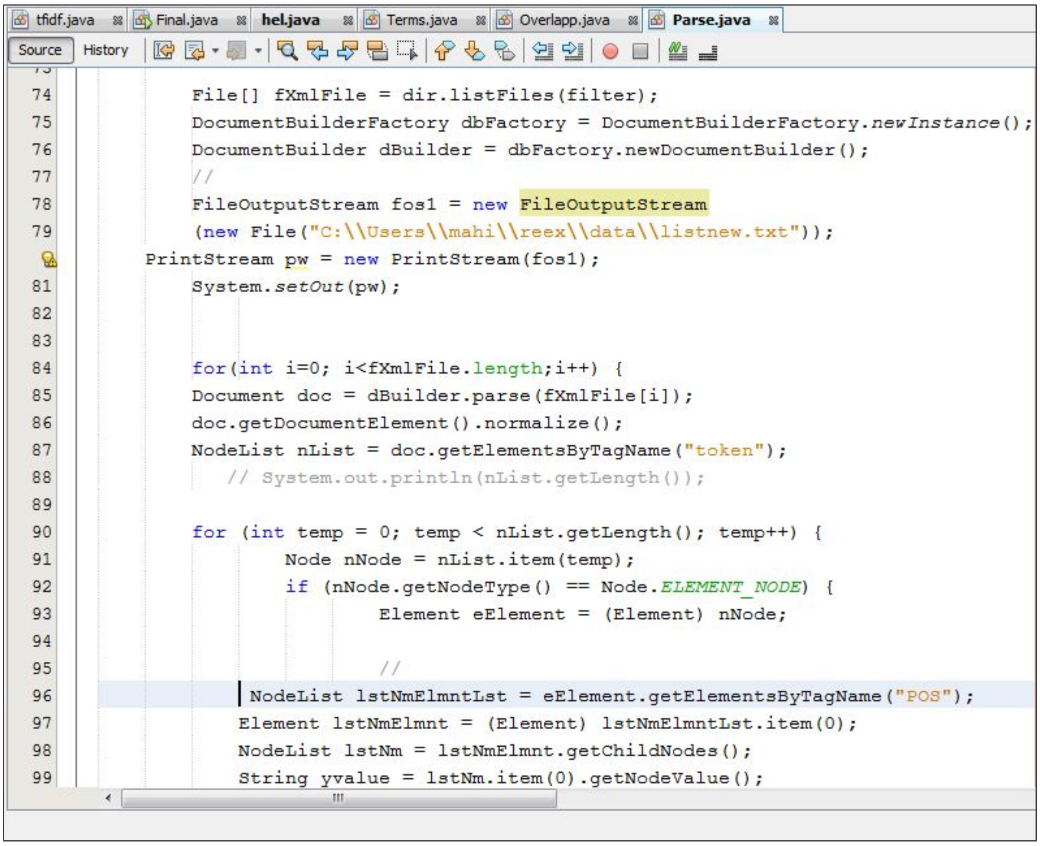
Tab. 5.2: Accuracy measures of each step.

Development Language	Java version 8 update 40
IDE	Eclipse and Weka
Libraries	StanfordCoreNLP, Lucene and Weka

The preprocessing step is further sub divided into three steps that are: (i) Tokenization (ii) Stop word removal and (iii) lemmatization Figure 5.1 shows the sub-code to get a processed dataset from a raw corpus dataset.

Getting processed dataset involves the following steps:

- Each Document in the Corpus is considered as a separate unit.



```
    File[] fXmlFile = dir.listFiles(filter);
    DocumentBuilderFactory dbFactory = DocumentBuilderFactory.newInstance();
    DocumentBuilder dBuilder = dbFactory.newDocumentBuilder();
    //
    FileOutputStream fos1 = new FileOutputStream
    (new File("C:\\Users\\mahi\\reex\\data\\listnew.txt"));
PrintStream pw = new PrintStream(fos1);
    System.setOut(pw);


    for(int i=0; i<fXmlFile.length;i++) {
    Document doc = dBuilder.parse(fXmlFile[i]);
    doc.getDocumentElement().normalize();
    NodeList nList = doc.getElementsByTagName("token");
       // System.out.println(nList.getLength());

    for (int temp = 0; temp < nList.getLength(); temp++) {
            Node nNode = nList.item(temp);
            if (nNode.getNodeType() == Node.ELEMENT_NODE) {
                    Element eElement = (Element) nNode;

                    //
         NodeList lstNmElmntLst = eElement.getElementsByTagName("POS");
        Element lstNmElmnt = (Element) lstNmElmntLst.item(0);
        NodeList lstNm = lstNmElmnt.getChildNodes();
        String yvalue = lstNm.item(0).getNodeValue();
```

Fig. 5.1: Parsing Document

- StanfordCoreNLP library is used to annotate each document by selecting appropriate annotator, result of which is the separate XML file for each document containing annotator.

- XML files are parsed and extract only useful POS tags that are required.

- StanfordCoreNlp also lemmatizes each word in the document during annotation.

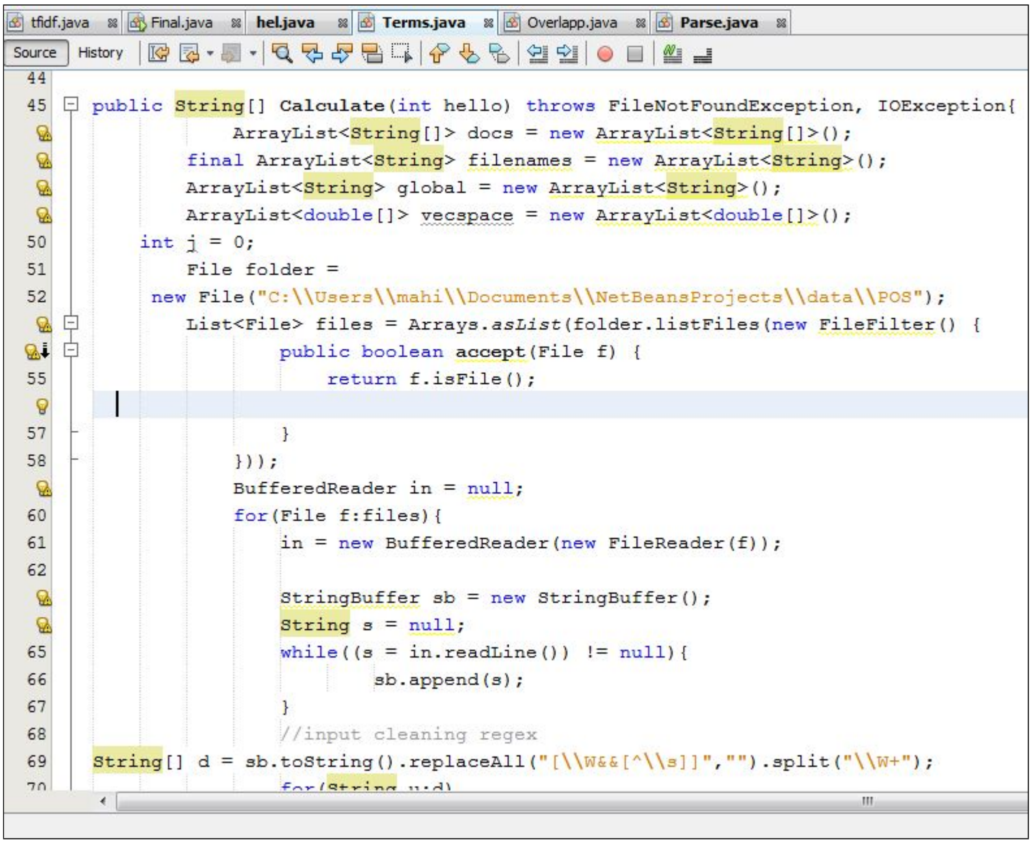
- Lucene stop word list is used to remove stop words during parsing.

With these five steps raw corpus is all processed.

Extracting Top Terms from the processed document need finding TF_IDF for each word in the corpus against each document. Figure 5.2 shows the sub-code for top terms.

- Hash map is used to create index containing all words in the corpus. TF_IDF weights are computed and saved as a value against each word in the index.

- Term-Document matrix (tdm) is created by populating matrix with TF_IDF weights.



```
public String[] Calculate(int hello) throws FileNotFoundException, IOException{
                ArrayList<String[]> docs = new ArrayList<String[]>();
            final ArrayList<String> filenames = new ArrayList<String>();
            ArrayList<String> global = new ArrayList<String>();
            ArrayList<double[]> vecspace = new ArrayList<double[]>();
        int j = 0;
            File folder =
       new File("C:\\Users\\mahi\\Documents\\NetBeansProjects\\data\\POS");
            List<File> files = Arrays.asList(folder.listFiles(new FileFilter() {
                    public boolean accept(File f) {
                        return f.isFile();

                    }
                }));
                BufferedReader in = null;
                for(File f:files){
                    in = new BufferedReader(new FileReader(f));

                    StringBuffer sb = new StringBuffer();
                    String s = null;
                    while((s = in.readLine()) != null){
                            sb.append(s);
                    }
                    //input cleaning regex
String[] d = sb.toString().replaceAll("[\\W&&[^\\s]]","").split("\\W+");
                    for(String u:d)
```

Fig. 5.2: Finding Top Terms

- Top terms are extracted by comparing each termâĂŹs weight in the tdm with the minimum threshold provided by the investigator.

The second phase is all about formulating and Expanding Subject vector.

Initial Subject Vector is created by investigators by Selecting suitable terms from the list of top terms given to them using GUI interface (figure 5.3) of the application.

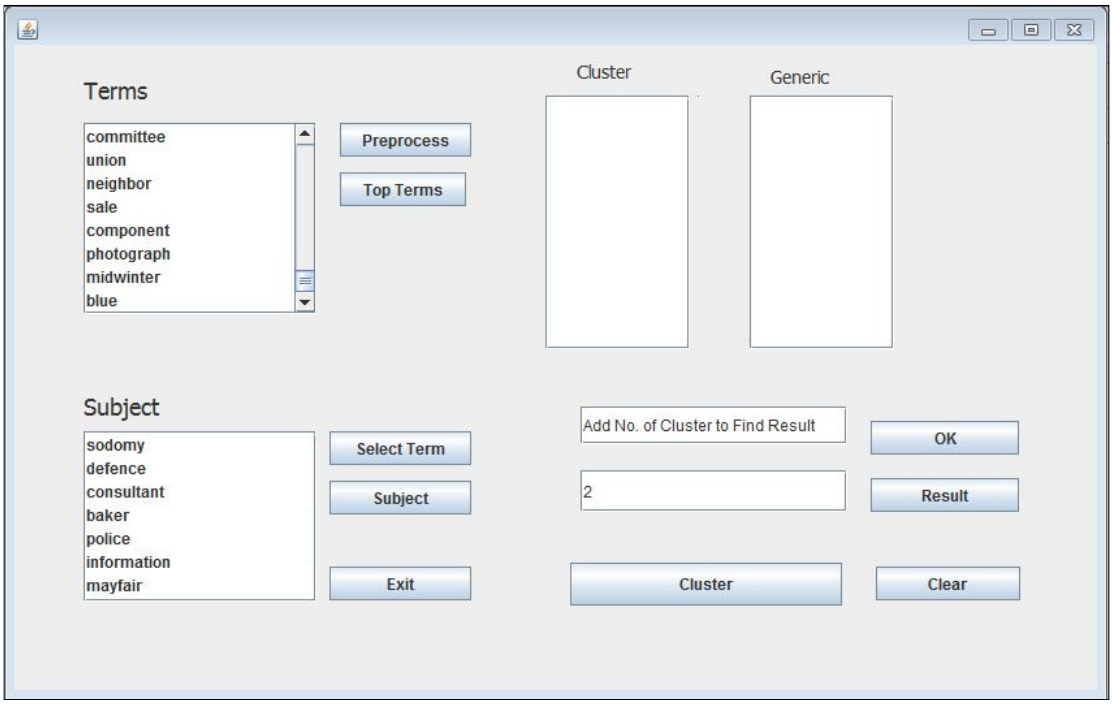
Word Net is further used to expand the list of terms selected by investigators. Sub code for this phase is shown in figure 5.4.

Third and fourth phase is all about clustering Documents using expanded subject vector. Online Source code for Overlap clustering is used for the said procedure(fig 5.5).

Fourth Phase is about repeating all the previous three phases for the generic cluster produced as the result of phase three clustering. Figure 5.6 shows sub-code for this phase.

5.1 Dataset Specifications

In order to conduct our experiment, we used dataset from two different sources to evaluate and compare our system with the previously available system. The description of each



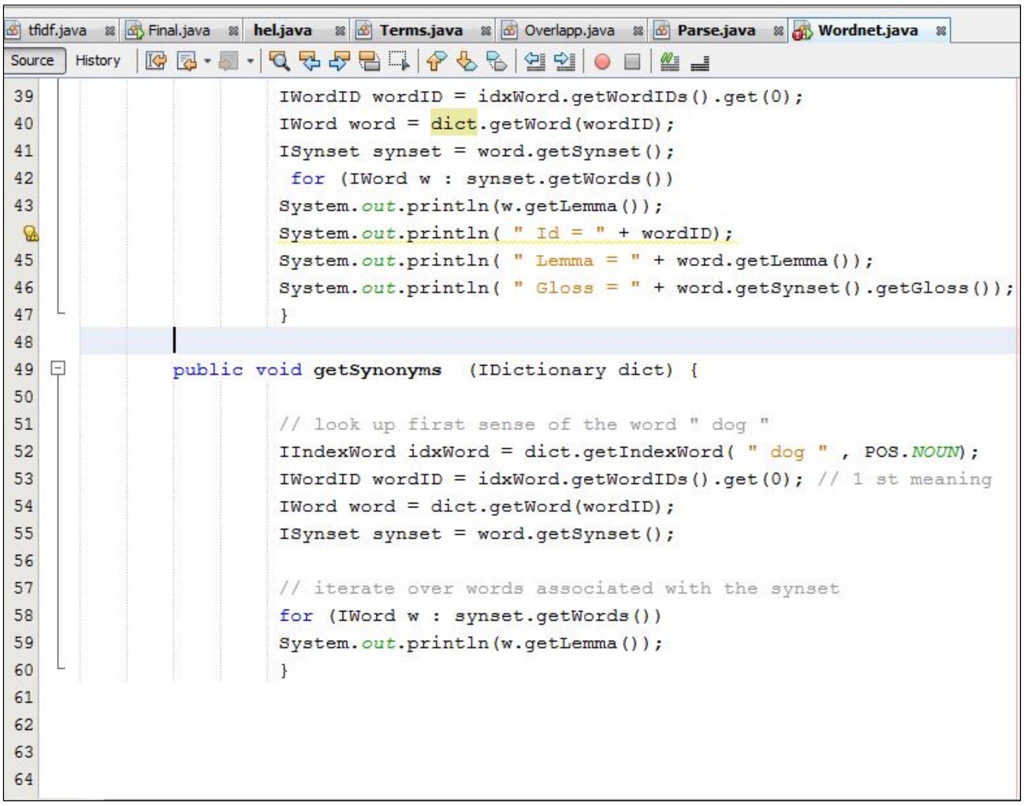
committee
union
neighbor
sale
component
photograph
midwinter
blue

sodomy
defence
consultant
baker
police
information
mayfair

Add No. of Cluster to Find Result

2

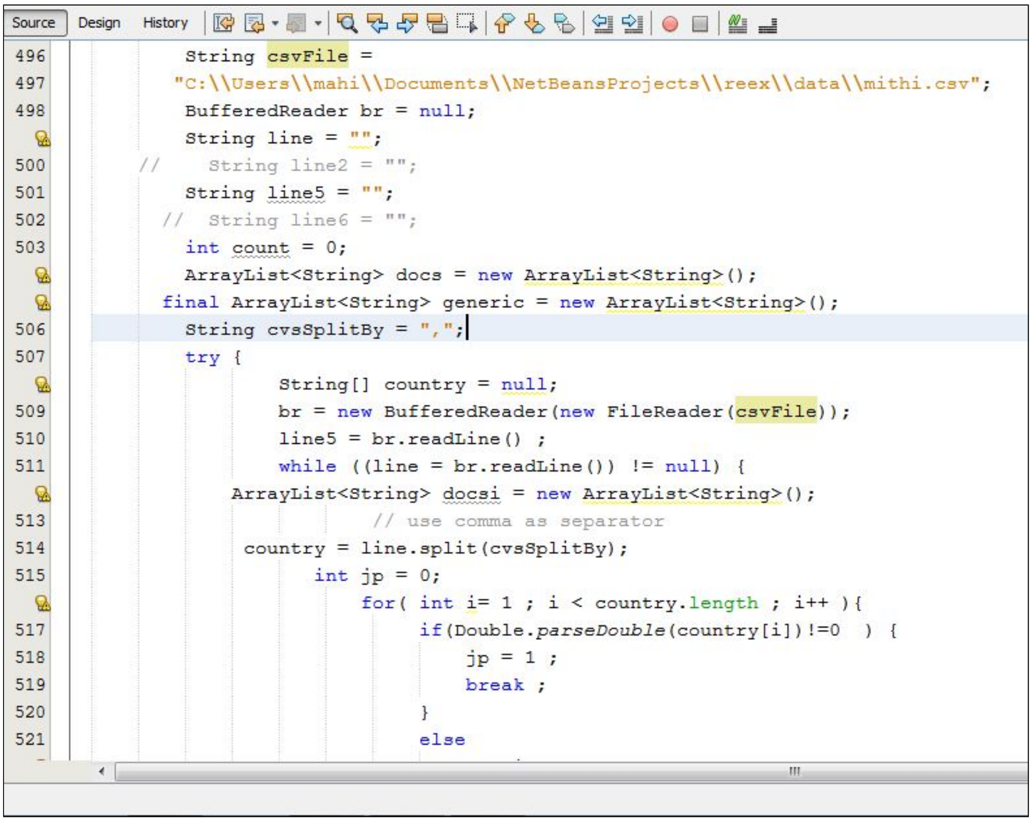
Fig. 5.3: Subject Formulation



```
                IWordID wordID = idxWord.getWordIDs().get(0);
                IWord word = dict.getWord(wordID);
                ISynset synset = word.getSynset();
                 for (IWord w : synset.getWords())
                System.out.println(w.getLemma());
                System.out.println( " Id = " + wordID);
                System.out.println( " Lemma = " + word.getLemma());
                System.out.println( " Gloss = " + word.getSynset().getGloss());
                }

        public void getSynonyms  (IDictionary dict) {

                // look up first sense of the word " dog "
                IIndexWord idxWord = dict.getIndexWord( " dog " , POS.NOUN);
                IWordID wordID = idxWord.getWordIDs().get(0); // 1 st meaning
                IWord word = dict.getWord(wordID);
                ISynset synset = word.getSynset();

                // iterate over words associated with the synset
                for (IWord w : synset.getWords())
                System.out.println(w.getLemma());
                }
```

Fig. 5.4: Finding Synonym from wordnet

data source is specified in the following subsections.



```
            String csvFile =
          "C:\\Users\\mahi\\Documents\\NetBeansProjects\\reex\\data\\mithi.csv";
            BufferedReader br = null;
            String line = "";
        //    String line2 = "";
            String line5 = "";
          //  String line6 = "";
            int count = 0;
            ArrayList<String> docs = new ArrayList<String>();
          final ArrayList<String> generic = new ArrayList<String>();
            String cvsSplitBy = ",";
            try {
                    String[] country = null;
                    br = new BufferedReader(new FileReader(csvFile));
                    line5 = br.readLine() ;
                    while ((line = br.readLine()) != null) {
                ArrayList<String> docsi = new ArrayList<String>();
                                // use comma as separator
                 country = line.split(cvsSplitBy);
                          int jp = 0;
                              for( int i= 1 ; i < country.length ; i++ ){
                                   if(Double.parseDouble(country[i])!=0  ) {
                                       jp = 1 ;
                                       break ;
                                   }
                                   else
```

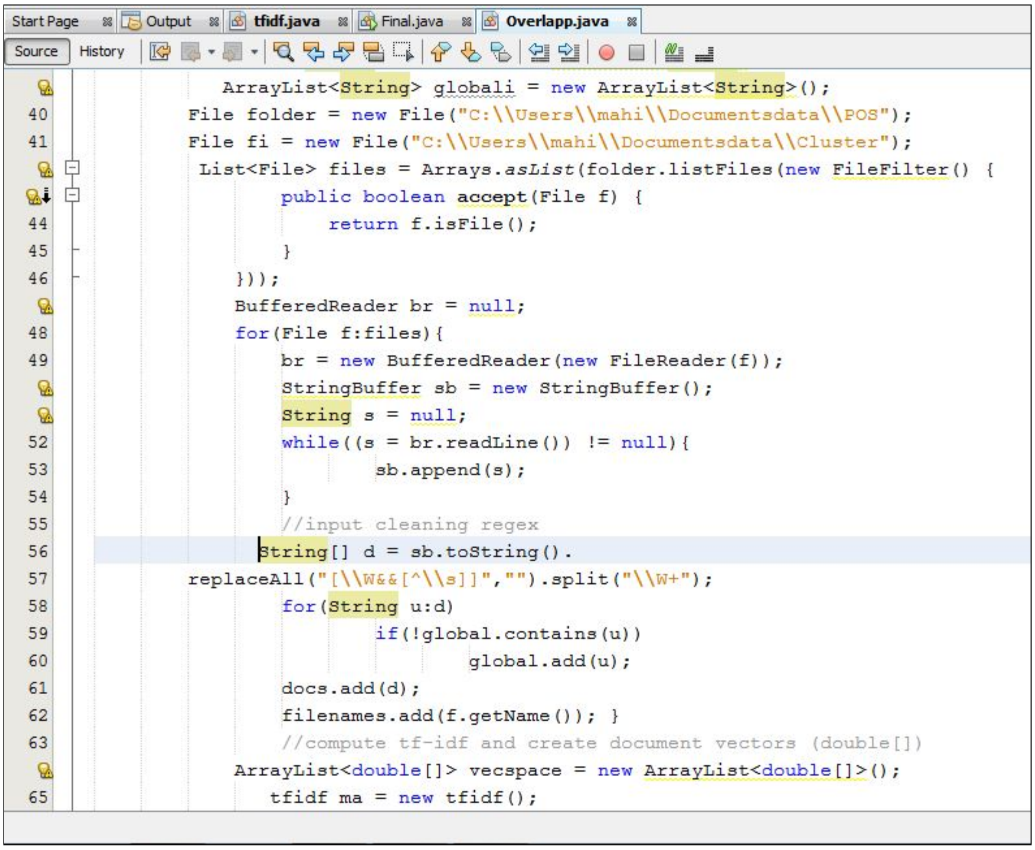
Fig. 5.5: Overlap clustering the document

5.1.1 Online Source

In order to test and evaluate our system, we collected a dataset of different newsgroup from archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups. The 20 newsgroup data set is a collection of 20,000 news documents arranged in 20 different newsgroups each belongs to different topic. We randomly choose 200 documents, 10 documents belongs to each 20 different newsgroups. As the documents are already arranged in different topics, it is easier to evaluate our semantic clustering system.

5.1.2 Past Paper Source

Our second dataset we got to validate our proposed system is the same dataset which the researchers Dagher, G. G. & Fung, B. C used in their experiment [10] . The dataset consists in total of 120 documents out of which 90 documents belongs to three different classes namely drugs, hacking and rape. Remaining 30 documents are overlapping documents belong to more than one topic (class).



```
                ArrayList<String> globali = new ArrayList<String>();
            File folder = new File("C:\\Users\\mahi\\Documentsdata\\POS");
            File fi = new File("C:\\Users\\mahi\\Documentsdata\\Cluster");
             List<File> files = Arrays.asList(folder.listFiles(new FileFilter() {
                     public boolean accept(File f) {
                         return f.isFile();
                     }
                 }));
                 BufferedReader br = null;
                 for(File f:files){
                     br = new BufferedReader(new FileReader(f));
                     StringBuffer sb = new StringBuffer();
                     String s = null;
                     while((s = br.readLine()) != null){
                             sb.append(s);
                     }
                     //input cleaning regex
                   String[] d = sb.toString().
            replaceAll("[\\W&&[^\\s]]","").split("\\W+");
                     for(String u:d)
                             if(!global.contains(u))
                                     global.add(u);
                     docs.add(d);
                     filenames.add(f.getName()); }
                     //compute tf-idf and create document vectors (double[])
                 ArrayList<double[]> vecspace = new ArrayList<double[]>();
                    tfidf ma = new tfidf();
```

Fig. 5.6: Subject Formulation

5.2 Evaluation Metrics

We have used F-measure [ref] as our evaluation metric to measure our system performance and accuracy. F-measure is the harmonic mean of precision and recall. Formula for F-measure is

$$Precision = 2 \ast \frac{Precision \ast Recall}{Precision + Recall}$$

Before going in detail of using these metrics, it is important to know the terms used in this evaluation metric, for example TP (True Positive), FP (False Positive), TN (True Negative) and FN (False Negative) as shown in Confusion Matrix (table 5.3).

Tab. 5.3: Confusion Matrix.

	Same Cluster	Different Cluster
Same Class	TP	FN
Different Class	FP	TN

- TP: When expected and actual value of the cluster both are true.

- FN: We predicted No, when the value for clustering is actually true.
- FP: We predicted Yes, when the value for clustering is actually false.
- TN: Both actual and predicted value is negative.

Recall: Recall is how accurately each document in the dataset is clusetered into their relevant cluster. It is also called Sensitivity. Formula for the recall is

$$Recall = \frac{TP}{P}$$

Precision: Precision is measured as the fraction of Documents correctly placed in the same cluster. In other words, it calculate how many documents that system identified are correct. Formula for precision is

$$Precision = \frac{TP}{TP + FP}$$

5.3 Results and Analysis

The proposed solution is implemented on a dataset for crime investigation. The data is grouped into some clusters, given as Rape, Hacking, Drug and Generic. Then the generic cluster is broken into these three crime categories again and again until no document left in generic cluster. Confusion metrics have been made on each type of cluster in each case. Then the accuracy is calculated for both approaches (leaving a generic cluster as it is, or breaking the generic cluster again and again) on the basis of Recall, Precision and F-measure. This lengthy process is divided into set of process in each case.
Documents are clustered using overlapping algorithm by taking different values of δ for example 0.2, 0.3 and 0.4. Here the clustering algorithm uses δ value of 0.2.

5.3.1 Step 1, Semantic document Clustering

In this step the documents with most frequent term have been extracted (through WordNet) and formed a cluster (through documents clustering algorithm), as shown in table. First the documents having most frequent term âĂIJrapeâĂİ are grouped into the rape cluster.

Calculation of Precision, Recall and Frequency measure of Rape cluster is as follows:

$$Recall = 28/30 = 0.93$$

Tab. 5.4: Documents part of RAPE cluster and some overlapping documents

Cluster	Documents	Overlap Documents	False Positive
Rape	rape_1, rape_11, rape_13, rape_14, rape_15, rape_16, rape_17, rape_18, rape_19, rape_2, rape_20, rape_21, rape_22, rape_24, rape_25, rape_26, rape_27, rape_28, rape_29, rape_3, rape_30, rape_31, rape_4, rape_5, rape_6, rape_7, rape_8, rape_9	hacking_21	hacking24, hacking29

Tab. 5.5: Confusion Matrix for Rape Cluster

True Positive(28)	False Negative(2)
False Positive(2)	

$$Precision = 28/30 = 0.93$$

$$F - measure = (2 * 0.93 * 0.93)/(0.93 + 0.93) = 0.92$$

Then, the documents having most frequent term âĂIJHackingâĂİ are grouped into the hacking cluster.

Calculation of Precision, Recall and Frequency measure are as follows:

$$Recall = 20/30 = 0.6$$

$$Precision = 20/25 = 0.8$$

$$F - measure = (2 * 0.6 * 0.8)/0.6 + 0.8 = 0.68$$

The documents having âĂIJdrugsâĂİ like terminologies are grouped into the drug cluster.

Tab. 5.6: Documents part of hacking cluster and some overlapping documents.

Cluster	Documents	Overlap Documents	False Positive
Hacking	hacking_1, hacking_11, hacking_12, hacking_15, hacking_16, hacking_2, hacking_20, hacking_21, hacking_22, hacking_23, hacking_25, hacking_26, hacking_27, hacking_28, hacking_3, hacking_4, hacking_5, hacking_6, hacking_8, hacking_9	drugs_11, drugs_16, drugs_17,rape_13, rape_15, rape_16, rape_17, rape_24, rape_25, rape_26, rape_27, rape_28, rape_29, rape_30, rape_31	rape_10,drugs_18, drugs_24, drugs_25,rape_23

Tab. 5.7: Confusion Matrix for Hacking Cluster

True Positive(20)	False Negative(10)
False Positive(5)	

Calculation of its Precision, Recall and Frequency measure are as follows:

$$Recall = 16/30 = 0.53$$

$$Precision = 16/18 = 0.88$$

$$F - measure = (2 * 0.53 * 0.88)/0.53 + 0.88 = 0.65$$

After grouping data in the form of cluster on the basis of those three crimes, the remaining documents would be a part of this Generic cluster.

The average accuracy of the crime clusters formed in this step is taken by adding their corresponding f-measures and no.of subjects.

$$AverageAccuracy = (F-measure\_rape+F-measure\_Drug+F-measure\_hacking)/3$$

Tab. 5.8: Documents part of Drug cluster and some overlapping documents.

Cluster	Documents	Overlap Documents	False Positive
Drug	drugs_11, drugs_12, drugs_16, drugs_17, drugs_2, drugs_20, drugs_27, drugs_28, drugs_30, drugs_4, drugs_9, drugs_7, drugs_8, drugs_23, drugs_29, drugs_3	hacking_2, hacking_23, hacking_3, rape_1, rape_15, rape_19, rape_20, rape_3, rape_30, rape_31, rape_6, rape_9	hacking_14, hacking_29

Tab. 5.9: Confusion Matrix for Drug Cluster

True Positive(16)	False Negative(14)
False Positive(2)	

Tab. 5.10: Documents part of Generic cluster.

Generic	drugs_1, drugs_10, drugs_13, drugs_14, drugs_15, drugs_19, drugs_21, drugs_22, drugs_26, drugs_5, drugs_6, hacking_10, hacking_13, hacking_17, hacking_18, hacking_19, hacking_30, hacking_7

$$= 0.92 + 0.68 + 0.65/3 = 0.75$$

After formulating clusters in the first step, Generic cluster is further re-clustered in the second step.

5.3.2 Step-2, Re-clustering of Generic Cluster

In this step the generic cluster is broken in two type of clusters, hacking and drug.

Calculation of its Precision, Recall and Frequency measure are as follows:

$$Recall = 27/30 = 0.9$$

Tab. 5.11: Clusters formed from Generic cluster

Drug	Hacking
drugs_1, drugs_10, drugs_13, drugs_14, drugs_15, drugs_19, drugs_21, drugs_22, drugs_26, drugs_5, drugs_6	hacking_10, hacking_13, hacking_17, hacking_18, hacking_19, hacking_30, hacking_7, drugs_26, drugs_6

Tab. 5.12: Confusion Matrix for Drug Cluster from Generic Cluster

True Positive(27)	False Negative(3)
False Positive(5)	

$$Precision = 27/32 = 0.84$$

$$F - measure = (2 * 0.9 * 0.84)/0.9 + 0.84 = 0.86$$

Tab. 5.13: Confusion Matrix for Hacking Cluster from Generic Cluster

True Positive(27)	False Negative(3)
False Positive(2)	

Calculation of its Precision, Recall and Frequency measure are as follows:

$$Recall = 27/30 = 0.9$$

$$Precision = 27/29 = 0.93$$

$$F - measure = (2 * 0.9 * 0.93)/0.9 + 0.93 = 0.91$$

The average accuracy of the crime clusters at step 2 is taken by adding the f-measure of rape cluster formed in step 1 with the corresponding f-measures of clusters formed in this step and no.of subjects.

$$AverageAccuracy = (F{-}measure\_rape{+}F{-}measure\_Drug{+}F{-}measure\_hacking)/3$$

$$= 0.92 + 0.86 + 0.91/3$$

$$= 0.9$$

It shows the impact of dividing the generic cluster again on accuracy. The accuracy of both approaches is shown in the table below. So it can be deduced that through our proposed approach of dividing the generic table, we can get higher accuracy.

Tab. 5.14: Accuracy measures of each step.

Accuracy at Step 1	Accuracy at Step 2
0.75	0.9

Similarly experiment is conducted by taking δ values of 0.3 and 0.4. The Results of these experiments at step 1 and at step 2 are shown in table 5.15 and 5.16 respectively.

Tab. 5.15: Accuracy measures at step 1

Minimum Similarity	Accuracy at Step 1
0.2	0.75
0.3	0.74
0.4	0.72

Tab. 5.16: Accuracy measures at step 2

Minimum Similarity	Accuracy at Step 2
0.2	0.9
0.3	0.81
0.4	0.80

5.4 Performance Evaluation

We evaluated the proposed system against δ values of 0.2, 0.3 and 0.4. We compute accuracy by finding Precision, Recall and F-measure for each δ value to evaluate the proposed system. rom the table 5.17, it can be seen that when re-clustering of generic clustering is performed in the step 2, accuracy is improved for all the three values of

Tab. 5.17: Accuracy measures at step 2

Minimum Similarity	Accuracy at Step 1	Accuracy at Step 2
0.2	0.75	0.9
0.3	0.74	0.81
0.4	0.72	0.80

δ. Hence proven that re-clustering of generic cluster by repeating the steps of proposed system improves the performance of the system.

5.5 Summary

This chapter provides implementation architecture of the system as well as the evaluation techniques used. Software and hardware specification is defined. Output of the proposed system is also briefly explained through screenshots. Dataset used for evaluation is explained briefly. The chapter also shows results on dataset provided by [10]. Comparison is provided against F-measure at both step 1 and step 2.

6. CONCLUSION AND FUTURE WORK

This chapter concludes the research work in the conclusion section. A brief description of the contributions and how the research work presented in this thesis can be extended for future work is also presented.

6.1 Conclusion

The primary purpose of research is to develop a DFI system motivated by using data mining techniques that is efficient and accurate. This system investigates crime documents by clustering in predefined subjects. In addition, generic cluster formed as a result of clustering is further clustered using the same semantic overlapping clustering algorithm. The proposed system uses preprocessing, NLP and clustering technique. The system is evaluated against dataset provided by Dagher, G. G. & Fung, B. C. The model provides more than accuracy above 85% . High accuracy is achieved because the proposed system uses lemmatization as it is more efficient to use with WordNet and clustering the generic cluster by going through iterations provides more information.

6.2 Contributions

Addition of subject suggestion module in the proposed system resulted in more accurate results when compared to previous techniques. Integration of lemmatization and further generic clustering in the system helped in making it more effective. This research will help the investigator to analyze documents in an efficient and timely manner.

6.3 Future Work

For future work, this research has opened new doors in DFI. First of all this research limits the terms in subject to nouns and verbs. However, it would be interesting to use adjective and adverb as part of terms in a subject. With increasing data size, dimensionality also increases hence, in future dimensionality reduction techniques such as Principal Component Analysis (PCA) and Outlier Detection techniques could be used. In order to further improve the system, it may use Wikipedia or other thesaurus to find synonym other than WordNet.

BIBLIOGRAPHY

[1] P. Berkhin *et al.*, “A survey of clustering data mining techniques.” *Grouping multidimensional data*, vol. 25, p. 71, 2006.

[2] Z. Huang, “A fast clustering algorithm to cluster very large categorical data sets in data mining.” *DMKD*, vol. 3, no. 8, pp. 34–39, 1997.

[3] A. Joshi and R. Kaur, “A review: Comparative study of various clustering techniques in data mining,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 3, 2013.

[4] N. Beebe and G. Dietrich, “A new process model for text string searching,” in *IFIP International Conference on Digital Forensics.* Springer, 2007, pp. 179–191.

[5] B. Carrier and E. H. Spafford, “An event-based digital forensic investigation framework,” in *Digital forensic research workshop*, 2004, pp. 11–13.

[6] G. Palmer *et al.*, “A road map for digital forensic research,” in *First Digital Forensic Research Workshop, Utica, New York*, 2001, pp. 27–30.

[7] M. Reith, C. Carr, and G. Gunsch, “An examination of digital forensic models,” *International Journal of Digital Evidence*, vol. 1, no. 3, pp. 1–12, 2002.

[8] G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[9] T. N. Dao and T. Simpson, “Measuring similarity between sentences,” *WordNet. Net, Tech. Rep*, 2005.

[10] G. G. Dagher and B. C. Fung, “Subject-based semantic document clustering for digital forensic investigations,” *Data & Knowledge Engineering*, vol. 86, pp. 224–241, 2013.

[11] T. Kalaikumaran, S. Karthik *et al.*, "Criminals and crime hotspot detection using data mining algorithms: clustering and classification," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 1, no. 10, pp. pp–225, 2012.

[12] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," *Computer*, vol. 37, no. 4, pp. 50–56, 2004.

[13] R. Adderley, M. Townsley, and J. Bond, "Use of data mining techniques to model crime scene investigator performance," *Knowledge-Based Systems*, vol. 20, no. 2, pp. 170–176, 2007.

[14] S. V. Nath, "Crime pattern detection using data mining," in *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on.* IEEE, 2006, pp. 41–44.

[15] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results," *Digital investigation*, vol. 4, pp. 49–54, 2007.

[16] P. Bide and R. Shedge, "Improved document clustering using k-means algorithm," in *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on.* IEEE, 2015, pp. 1–5.

[17] L. F. da Cruz Nassif and E. R. Hruschka, "Document clustering for forensic analysis: An approach for improving computer inspection," *IEEE transactions on information forensics and security*, vol. 8, no. 1, pp. 46–54, 2013.

[18] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, "Text clustering for digital forensics analysis," in *Computational Intelligence in Security for Information Systems.* Springer, 2009, pp. 29–36.

[19] D. B. Skillicorn and N. Vats, "Novel information discovery for intelligence and counterterrorism," *Decision Support Systems*, vol. 43, no. 4, pp. 1375–1382, 2007.

[20] G. Thilagavathi and J. Anitha, "Document clustering in forensic investigation by hybrid approach," *International Journal of Computer Applications*, vol. 91, no. 3, 2014.

[21] S. Mascarnes and J. Gomes, “Subject based clustering for digital forensic investigation with subject suggestion,” *International Journal of Computer Applications*, vol. 102, no. 11, 2014.

[22] G. Oatley, B. Ewart, and J. Zeleznikow, “Decision support systems for police: Lessons from the application of data mining techniques to ŞsoftŤ forensic evidence,” *Artificial Intelligence and Law*, vol. 14, no. 1-2, pp. 35–100, 2006.

[23] N. L. Beebe and L. Liu, “Clustering digital forensic string search output,” *Digital Investigation*, vol. 11, no. 4, pp. 314–322, 2014.

[24] N. L. Beebe, J. G. Clark, G. B. Dietrich, M. S. Ko, and D. Ko, “Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies,” *Decision Support Systems*, vol. 51, no. 4, pp. 732–744, 2011.

[25] C. Manning, P. Raghavan, and H. Schütze, “Introduction to information retrieval/christopher d,” 2009.

[26] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, “The stanford corenlp natural language processing toolkit.” in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[27] G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.