IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-2, Issue-4 E-ISSN: 2347-2693

Enhancing the Productivity of Digital Data Retrieval

S.Gowri[1*], G.S.Anandhamala[2] and G.Divya[3]

[1,3]Department of Information Technology, Sathyabama University, India*
[2] Department of Computer Science, St. Joseph's College of Engineering, India

www.ijcseonline.org

Received: 15/03/2014 Revised: 29/03/2014 Accepted: 21/04/2014 Published: 30/04/2014

Abstract— The escalate of crime rate around the global manifests the lag in the ongoing methodology in extraction or inspection of information retrieved and stored in cache from divers communication channels of the digital investigation system. The Principle objective for the development of this strategy, is to revamp the existing methodology and develop a crime relates information mining framework for extracting and examine pertinent information from stored data which incorporates emails, chat threads and any text messages to discover the criminal activity and solve the enigma with the help of the certainty concealed within the data. This strategy is achieved by the prominence of the three aspects initiated to the input data of textual evidences such as mails, text messages, chat threads, etc. the three aspects are, the separation of the body and header part of a textual corpuses consisting of all kinds of textual evidences of diverse communication channels and preprocessing of text , which is achieved using regular expressions of PHP script and stemming process respectively, in order to make the mining process of text more reliable for the forensic department of crime investigation. Further the searching technique used in this methodology which is constructed for the most highly effective and efficient retrieval of data. Although the technique followed in this methodology is an ancient technique of separation of body and header in a mail, the main aspect of this methodology is to focus over the efficiency in searching technique built for the purpose of effectiveness. A Clustering algorithm is used in this methodology to improvise the system. This algorithm has mainly three feature, it is the alternative form of the Reverse Factor algorithm, it uses bit-parallelism simulation of the suffix automaton of x^R and it is efficiency is high if the pattern length is not longer than the memory-word size of the machine. Using this kind of technique to improvise the existing system would bring about a methodical procedure, which would initiate in a highly efficacious searching system, evidenced by the time complexity, precision and recall value. The preprocessing of text, that is stemming is done for easier understanding and convince of the user to suspect over threads and text document. Integrating all the factors specified would assist in easier reviewing of the text messages.

Index Term— Textual evidences, Stemming, Information retrieval, Preprocessing.

I. INTRODUCTION

The Exchange and broadcast of information in the communication system is done through many ways over various channels. Every single transformation of message is done in some or the other ways such as mails transfers, chat messengers, social networks sites, online messaging sites and many more. Everyone uses any of the services for communication purpose. Crime based communications are also done with the help of same service providers but by utilization of anonymous ways of communication channels which also come under the digitization technology only. Crime rate increases by increase in illicit pursuits like hijacks, bankruptcy, bomb blasts, kidnapping and many infringements which are intended from various places communicated through various communication channels. For example, cyber predators and pedophiles instigate victims search in various proxy communication channels which includes even public chat rooms. Due to the technology present in the investigation system they are unable to handle large amount of data, predominantly cause of too much manual work. Taking the current scenario into consideration the communication rate has been increased to a large extent which becomes a strenuous job for the investigators to handle such an enormous unstructured data and perform the analyses[10] of wary data in it. In the forensic department the approaches and tool they follow and use respectively are the ancient methodologies which are not effective enough over such enormous data retrieved from various sessions which are to be handled by. The two main drawbacks of the current system followed in the forensic department are: The current approach followed[1] by the investigation system would be useful only for data which is well organized and structured but not applicable over the unstructured data like mails, where only the header part is been considered and not the message part, apart from that aspect the other aspect is that the system used are mainly been concentrated over the network level data like path followed by the packets in the network, IP address tracking and so on but not the message content.

Taking the above issues at the forensic department a new system of approach[1] is been followed in order to improvise the existing system. The process of monitoring the textual evidences for the forensic department by improvising the effectiveness in textual processing and retrieval of information[2] using the initiated procedure brings out the accuracy in maintenance of data and effective retrieval of the output. The segregation of the header and the body part of the text messages are done using regular expressions, here the header part of the messages are been directly stored into

Corresponding Author: S Gowri
Department of Information Technology, Sathyabama University, India

the database created and the body part of the message part is been saved into the database as a reference link, on click of the link would display the body of text in an text box separately which is been stored as a notepad file format, for the further purpose of preprocessing of text in-order to make it more reliable on message and focus over the conversations taking place. The Clustering algorithm used for the searching is the algorithm which fascinates the users of the introduced system. The most fascinating part in this algorithm is that, the searching time and the process time. The searching time is approximate of 11-12 milliseconds and the processing time is approximate of 2-3 milliseconds. The most enormous problem in the world is big data. The establishment of this system brings upon an effective retrieval of data from any huge datasets. This system is carried out considering another aspect along with the searching technique that is the separation of header and the body part of a textual script. This technique is implemented using the regular expressions in scripting language. These Regular expressions are mainly employed for complex string manipulation. PHP implements the POSIX extended regular expressions as elucidated by POSIX 1003.2. The preprocessing of text is done using the Java programming, Here only the stemming in text mining process is been implemented.

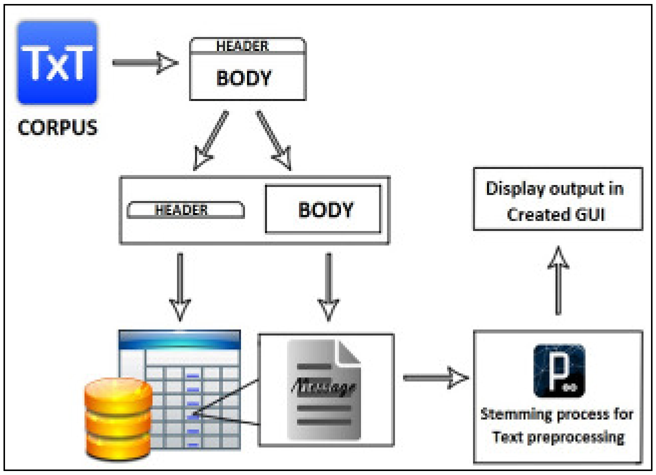

TxT
CORPUS
HEADER
BODY
HEADER
BODY
Display output in Created GUI
Message
Stemming process for Text preprocessing

Fig 1. System Architecture

II. OVERVIEW OF SYSTEM ARCHITECTURE

The overall system overview is set in a user understandable manner which could be a user friendly and very efficient system. Initially the text corpus is loaded into the created folder, this loaded mails are then been segmented into header and body of all kinds of textual documents. The segments are then arranges in such a way that the header part is stored directly into the table with the allocated fields and the message part is stored with its references in the table. The message part is then been viewed on click of the reference link in the table.

The message segment of the textually evidenced document is then been sent to the text preprocessing unit of an algorithm for stemming the text for a better understanding to an individual.

III. RESULT & EXPERIMENTAL ANALYSIS

A. Message

Specifically only one header is present in each message, in the form of fields where a name followed by value is given.

A separate field is set for each line of text beginning with a printable character in the header. The field name starts with a character and a separator character “:” is set to specify the end of field name, where a value is then followed after the end separator. If spaces or tabs are given as the first character at the value side then the value is continued onto subsequent lines. The 7-bit ASCII characters are only allowed for the field’s name and values, where the non ASCII values are represented using MIME encoded words.

The Fields of Email header are:

- fname→ File absolute location
- mid→ Mail Document Identifier
- subject→ Subject of the Email Document
- text→ Body content of the Document
- sender→Sender of the email
- rcvr→ Receiver(s) of the email
- date→ Date of the email
- mv→ Mime-Version
- ct→ Content Type
- cte→ Content Transfer Encoding
- cc→ cc contents of the email
- bcc→ bcc contents of the email
- xfrom→ sender name (how it has been saved in the contacts)
- xto→ receiver name (how it has been saved in the contacts)
- xcc→cc contents(how it has been saved in the contacts)
- xbcc→ bcc contents(how it has been saved in the contacts)
- xfolder→The folder which contains the email
- xorigin→ The Person Name (Owner of the current mail box)
- xfilename→ the name with which the email is stored
- timestamp →The time stamp of the email in the form (YYYYMMDDhhmmss)

Field names are case sensitive. For example “Subject” and “subject” will not be treated as same. The default data type of the email fields are “text”. If the field name of a term is not mentioned, then it will search for that term in the default field, that is text, which means “terrorist” and “text: terrorist” are equivalent.

B. *Segementation of body and header from text documents*

The main reason for the segmentation of the header and the body of a textual document is to give a clear analysis over both the address and the message information, so as to find the accurate over the retrieval. The head part is stored in the table with the respective columns and the body part is stored in the note pad with its reference in the table.

a) PHP is an open source language for assembling dynamic web pages. PHP possesses three sets of functions that permit to work with regular expressions.

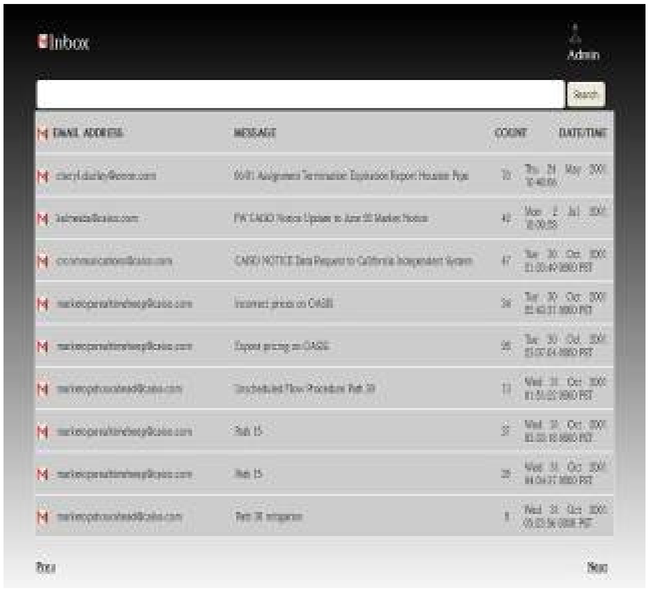
b) The most predominant set of regex functions is preg. PCRE library (Perl-Compatible Regular Expressions) is wrapped around with these regex functions. PHP incorporates PCRE by default as of PHP 4.2.0.

c) The former set of regex functions are the ones which start with ereg. The functions utilize the POSIX Extended Regular Expressions, which is similar to the traditional UNIX egrep command. These functions are predominant for backward affinity with PHP 3.

d) The last set is a different from the ereg set, "multibyte" prefix is mb_ to the function names. On the other hand ereg handles the regex and subject string as a sequence of 8-bit characters, mb_ereg can be implemented with multi-byte characters from diverse code pages. For regex to handle Far East characters as independent characters, mb_ereg functions are been implemented, or the /u modifier with preg functions.

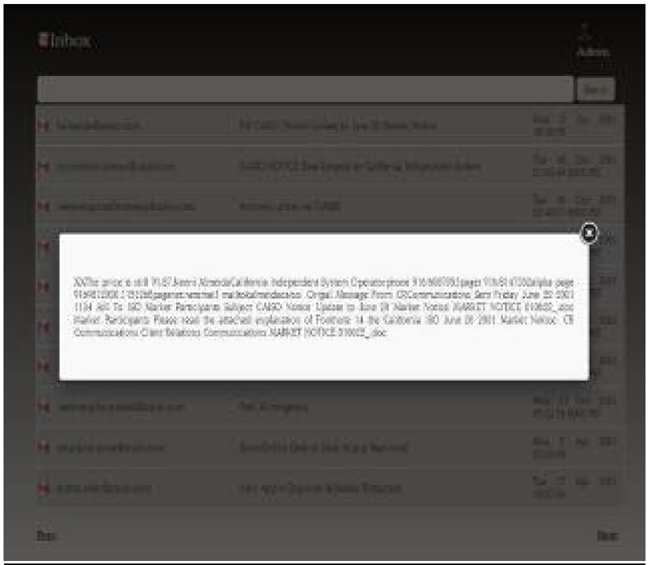
The standard pattern[4] defined for the separation using the regular expressions is:

('Return-Path', 'X-Original-To', 'Delivered-To', 'Received',' In-Reply-To', 'To', 'Message-Id', 'Date', 'From', 'Subject', 'Bcc', 'Cc', 'Precedence', 'Reference', 'Sender', 'Archived')



EMAIL ADDRESS	MESSAGE	COUNT	DATE/TIME
cheryl.dunley@enron.com	06/01 Assignment Termination Explanation Report Houston Pipe	70	Thu 24 May 2001 10:40:06
[illegible]@caiso.com	FW CAISO Notice Update to June 28 Market Notice	42	Mon 2 Jul 2001 10:30:58
crcommunications@caiso.com	CAISO NOTICE Data Request to California Independent System	47	Tue 30 Oct 2001 11:03:49 0800 PST
marketoperations[illegible]@caiso.com	Incorrect prices on OASIS	58	Tue 30 Oct 2001 05:40:31 0800 PST
marketoperations[illegible]@caiso.com	Expost pricing on OASIS	56	Tue 30 Oct 2001 05:07:04 0800 PST
marketopshourahead@caiso.com	Unscheduled Flow Procedure Path 30	13	Wed 31 Oct 2001 01:51:03 0800 PST
marketoperations[illegible]@caiso.com	Path 15	37	Wed 31 Oct 2001 [illegible] 0800 PST
marketoperations[illegible]@caiso.com	Path 15	28	Wed 31 Oct 2001 04:04:37 0800 PST
marketopshourahead@caiso.com	Path 26 mitigation	8	Wed 31 Oct 2001 05:03:56 0800 PST

Fig 2.1. PHP framework for separated header part



XXThe price is still [illegible] California Independent System Operator phone [illegible] Original Message From CRCommunications Sent Friday June 22 2001 11:54 AM To ISO Market Participants Subject CAISO Notice Update to June 28 Market Notice MARKET NOTICE 010628_doc Market Participants Please read the attached explanation of Footnote 14 the California ISO June 28 2001 Market Notice CR Communications Client Relations Communications MARKET NOTICE 010628_doc

Fig 2.2. PHP framework for separated body part

C. *Stemming of Segmented Text Document*

Stemming of the text document is mainly attained cause of two significant reasons, one for making the search of keyword more efficacious and the other for easier understanding of the document when manually inspected. Reducing a word to its base is called Stemming (or stem). For example, the words 'studying', 'studied' and 'studies' all have the stem 'study'[5]. A word, or a group of words, and generates the stem, or a group of the stems, of the input.

Stemming is convenient when any kind of text-analysis is done. Considering the contents of a text, the various times of verbs and the enormous endings for singular and plural, make it strenuous to discern the significance of particular words inside the text, when each word is treated as it is.
For example the text "I speak about humanity, after I had spoke on humanity for a speech. Currently I'm speaking about Human Rights. Next year I'll speak on women empowerment. But I made my favorite speeches when I was young."

Now, to a human, its known main consideration in the text is 'speech'. But when a program that does simple text analysis is executed over it, the conclusion of the program would be a text about human, since the only word that occurs more than once is 'human' (apart from words like 'i', 'a', etc. which are filtered out usually).

Analyzing the text before Stemming the text documents, that is all the words being replaced with their stems, the program will exactly give the output saying that the text is about speaking, because after stemming, the appearance of the word 'speak' will be four times, because 'speak', 'spoke ', 'speaking' and 'speak' have been rewritten by 'speak'.

Algorithm wise the stemming is a problem, because of the different rules and notable cases in the English language. But this algorithm for stemming[3] solves the problem of difficulty in utilization.

Coding process:

Initially a library has to be imported for text based manipulations; here we use JWNL (Java WordNet Library) Now we can create a class for the stemming algorithm. The class should contain these members:

Unfortunately, WordNet recovery of word information is not so quick, so usage of a HashMap to store stemmed form of the words, so stemming a word a more than once will be less in cost than a constant-time hashmap lookup.

To establish the connection with WordNet database initializing with file input stream properties[9]:

a) First creation of hashmap is done, then initializing of JWNL library and finally catch and report any kind of exceptions this might occur.

b) The JWNL is initialized in the try block with xml file properties which were copied into the folder of the application. Then a Dictionary and a MorphologicalProcessor is been generated, were actual stemming is required. After this sets the size of JWNL's internal cache to 10000, when then comments the outline. Eventually, an IsInitialized flag is set to true, which would prevent utilization of dictionary by the stemming functions.

c) The actual word stemming method: Initialization of JWNL. If not, return the input word. Definition of an IndexWord is the data structure of the library. Using API verb, noun, adjective or adverb is considered. Since current stemming word[11] is unknown, all four kinds are used till the match is found, lookupBaseForm() is used till a non-null return value. Now from the IndexWords the word which is stemmed is to be got, where IndexWord.getLemma().toString() is been used for this process.

The structure imitated for a word is:
[Prefix] Root { [Infix] Root } [Suffix]

Example: selection
selections
selective → select
selected
selecting

The process is a step by step process of eliminating words with the help of set of character matrixes which contains all the suffixes and will be categorized and sent into the loops and stemmed accordingly.

The algorithm parses the through all the text documents in the corpus and will follow the flow shown in the Fig 3. were the indexing will crop the word of after the root by either adding a space or a punctuation mark to show the end of word and finally the stemmed output is shown as the output snapshot in the Fig 4.

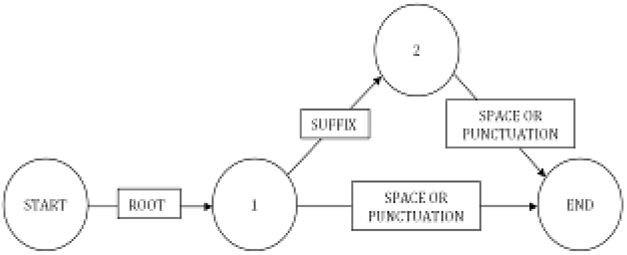


Fig 3. Flow of Stemming Approach

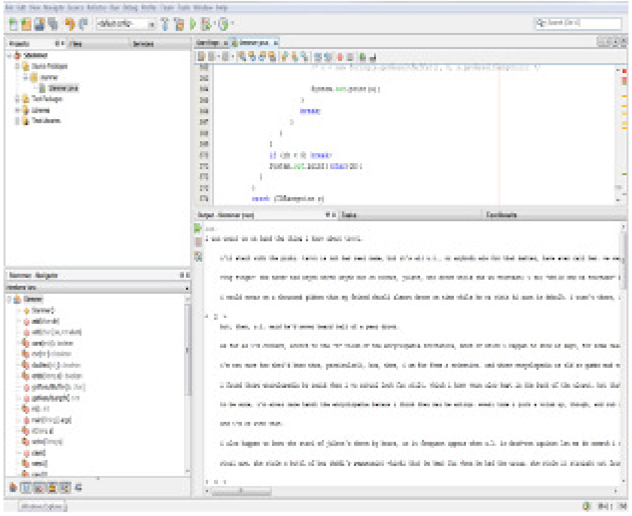


Fig 4. Stemming Ouput

D. The Culstering Algorithm for seacrching technique

The searching is done by the keyword as an input from the user in order to filter the mails and display accordingly which is then followed by the sort process to make it clear the content inside the message of the mail. The sort is done using the word_count variable set in the searching algorithm incrementing the value of count on found and using that count variable the sorting is done accordingly and displayed over the output screen .

1) Main features

a) This algorithm is the alternative form of the Reverse Factor algorithm.

b) Bit-parallelism simulation is employed of the suffix automaton of xR.

c) If the pattern length is not extended than the machine’s memory-word size, the algorithm is exceptionally effective.

2) Description

- Using the bit parallelism[8] this algorithm simulates the BDM algorithm.
- *Properties:*
  - The search is performed in a window which is made to move forward incrementally without looping.

  - The window shift occurs when there is no active state, which is similar to that of BDN, since no looping process is done.
- If d_m=1the match of string is returned.
- The match of longest prefix w in S, the variable would last correspondence.
- The Search status updating formula is:

  Dj=(Dj-1 AND B[Sj]) << 1
- Explanation of the update formula
  - Only an active machine can receive a new symbol, no input symbol can re-activate any machine
  - The next input symbol determines on which active machine it will be possible to try to process it
    - The bit mask determines which machine can process the input symbol (those having the corresponding 1 in the bit mask, i.e. bi=1)
    - To remain active, a machine must previously be active (di=1) and the bit mask for that machine must be 1 (therefore AND in the formula)
  - The fact that the backward scan of the window reflect the left shit by one
- Compared to Shit-And and Shift-Or the bit mask has reverse order, since the search is backwards.

Example: if *w*=bbaac, *B*[a]=00110, *B*[b]=11000, *B*[c]=00001

Algorithm: Text search

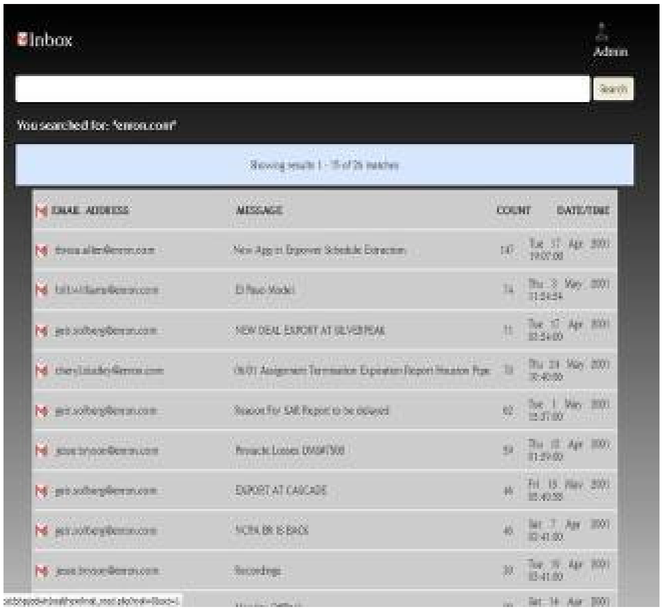
```
Procedure CA(*x, m, *y, n)
j: {The position of the string present}
s: {Input}
    Begin
    if m > WORD_SIZE
    error("Culturing");
    /* Pre processing */
    Memset to(B,0,ASIZE*sizeof(int));
    S←1;
    for : i← m to 0 decrement
        Begin
        B[x[i]] |← s;
        s <<= 1;
        end
    /* Searching phase */
    J←0;
    while (j <= n-m)
      Begin
      I←m-1; last← m;
      d ← ~0;
      while (i>=0 && d!=0)
          Begin
          d &← B[y[j+i]];
          decrement i;
          if (d != 0)
              Begin
              if (i >= 0)
                last ← i+1;
              else
                OUTPUT(j);
              end
          d <<= 1;
          end
      j ← j+last;
      end
    end
```

Table B of size ASIZE in this clustering algorithm is initialized for individual character c, for which a bit mask is stored. If and only if x_i=c the mask in B_c is set. In a word d=d_{m-1} .. d_0 the search state is kept, where the machine word size is greater than or equal to the pattern length m.

If an only if x[m-i .. m-1-i+k]=y[j+m-k .. j+m-1] the bit d_i at iteration k is set. Variable d is set to 1^{m-1} at iteration 0. To update d follows d'= (d & B[y_j]) << 1 formula.

If and only if, after iteration m, it holds d_{m-1}=1, then there is a match.

The algorithm has emulated a prefix of the pattern in the present window position j, whenever d_{m-1}=1. The shift to the next position is given when the longest prefix is emulated.



You searched for: "enron.com"

Showing results 1 - 15 of 26 matches

EMAIL ADDRESS	MESSAGE	COUNT	DATE/TIME
drosa.allen@enron.com	New App to Expower Schedule Extraction	147	Tue 17 Apr 2001 19:07:00
bill.williams@enron.com	El Paso Model	74	Thu 3 May 2001 11:54:54
geir.solberg@enron.com	NEW DEAL EXPORT AT SILVERPEAK	11	Tue 17 Apr 2001 02:54:00
cheryl.dudley@enron.com	06/01 Assignment Termination Expiration Report Houston Pipe	13	Thu 24 May 2001 20:40:00
geir.solberg@enron.com	Reason For SAR Report to be delayed	62	Tue 1 May 2001 23:37:00
jesse.bryson@enron.com	Pinnacle Losses DWSF500	59	Thu 12 Apr 2001 01:29:00
geir.solberg@enron.com	EXPORT AT CASCADE	46	Fri 18 May 2001 05:40:00
geir.solberg@enron.com	NCPA BR IS BACK	46	Sat 7 Apr 2001 02:41:00
jesse.bryson@enron.com	Recordings	30	Tue 10 Apr 2001 02:41:00
[illegible]	[illegible]	[illegible]	Sat 14 Apr 2001

Fig 5. Search Result

3) *Precision and Recall*

To evaluate the searching technique strategies, the basic measures used are Precision and Recall[6]. There are 'N' numbers of records present in the corpus which are related to the search topic. The Records are presumed either as relevant or irrelevant. The set of relevant records may not be the set of records retrieved. The Fig 6. Graph shows the Comparison of the precision and recall values of this clustering algorithm to the reverse factoring algorithm which is kind of an alternative form shows the data retrieval which is comparatively higher in the percentage rate and more effective in the retrieval of relevant documents.

Precision: It is the ratio of number of relevant records to the number of irrelevant records and relevant records retrieved from the database on search is called precision[6], which is expressed in terms of percentage.
Recall: It is the ratio of the number of relevant records to that of the total number of relevant records retrieved in the database is called Recall[6], which is expressed in terms of percentage.

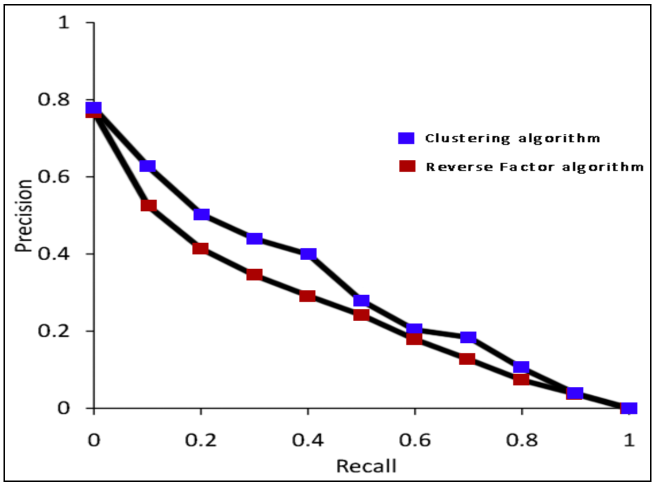


Fig 6. Precision and Recall graph for 50 Quires

IV. Comparison of Searching Techniques

The Fig 7. graph shows the time complexity[7] of various searching techniques, where the vertical axis represents the average time in milliseconds every algorithm requires to preprocess and parse the stemmed text document. Thus the algorithm with the least value of time is better. The proposed algorithm has the search time of approximately 11.2 milliseconds and the preprocessing time of approximately 2 milliseconds which would give a faster performance with an effective retrieval of relevant documents, as mentioned with the precision recall graph for the the algorithm.

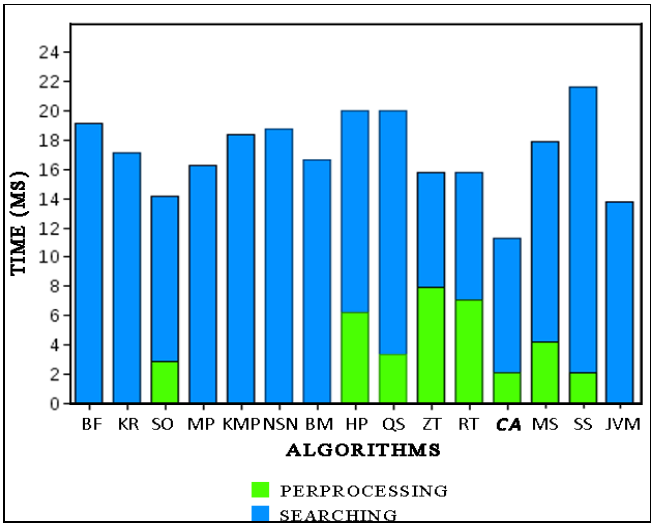


Fig 7. Comparison chart for Searching Algorithms

V. Conclusion

In this paper the performance of analyzing and mining (Stemming) over the collected data on textual documents like mails, message threads, chat threads etc., for which a system has been generated in which the segmentation of the header and body part of the corpuses and then the segmented body part is then manipulated using stemming procedure in java for a better viewing and searching process of keywords and phrases, the searching process used in this system is the newly introduced and implemented process. Integrating all these processes introduced and generated are been performed to find out the information related to crime in the forensic investigations in digital text retrieval. As per the work progress the test over mails are been done. Other then mails there are communication channels through which text has been passed in the later experiment consideration of all those communication is done. Furthermore a decision making model with more better performance on text documents needs to be worked on which able to perform auto detection of suspected text documents on the basis of training set and samples provided to it.

References

[1] S.Gowri; G.S.Anandha Mala; "Improving Intelligent IR Effectiveness in Forensic Analysis" Institution of Computer Science Informatics and Telecommunication Engineering 2012, Page(s): 451.

[2] Ms. Vandana Dhingra; Dr. Komal Kumar Bhatia; "Towards Intelligent Information Retrieval on Web" International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397Vol. 3 No. 4 Apr 2011

[3] Gabarro, S. "String Manipulations Revisited" Web Application Design and Implementation:Apache 2, PHP5, MySQL, JavaScript, and Linux/UNIX Digital Object Identifier: 10.1109/9780470083963.ch17 Page(s): 209- 216 Copyright Year: 2007.

[4] Minamide,Y. ; Sakuma,Y. ; Voronkov, A."Translating Regular Expression Matching into Transducers" Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2010 12th International Symposium on Digital Object Identifier: 10.1109/SYNASC.2010.50 Publication Year: 2010 , Page(s): 107- 115 Cited by: Papers (2).

[5] Bartoli, A. ; Davanzo, G. ; De Lorenzo, A. ; Medvet, E. ; Sorio, E. "Automatic Synthesis of Regular Expressions from Examples" Computer Volume:PP , Issue: 99 Digital Object Identifier: 10.1109/MC.2013.403 Publication Year: 2013 , Page(s): 1.

[6] Jesse Davis; Mark Goadrich; "The Relationship Between Precision-Recall and ROC Curves" Appearing in Proceedings of the 23rd international conference on Machine Learning, Pittsburg, PA, 2006.

[7] Suzumura, T. ; Trent, S. ; Tatsubori, M. ; Tozawa, A. ; Onodera, T. "Performance Comparison of Web Service Engines in PHP, Java and C" Web Services, 2008. ICWS '08. IEEE International Conference on Digital Object Identifier: 10.1109/ICWS.2008.71 Publication Year: 2008 , Page(s): 385- 392 Cited by: Papers (1).

[8] NAVARRO G., RAFFINOT M., 1998. "A Bit-Parallel Approach to Suffix Automata: Fast Extended String Matching", In Proceedings of the 9th Annual Symposium on

Combinatorial Pattern Matching, Lecture Notes in Computer Science 1448, Springer-Verlag, Berlin, 14-31.

[9] Nascimento, M.A. ; Da Cunha, A.C.R.,"An experiment stemming non-traditional text" String Processing and Information Retrieval: A South American Symposium, 1998. Proceedings Digital Object Identifier: 10.1109/SPIRE.1998.712985 Publication Year: 1998.

[10] Inikpi O. Ademu, Dr Chris O. Imafidon, Dr David S. Preston, "A New Approach of Digital Forensic Model for Digital Forensic Investigation" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.12, 2011.

[11] S.Gowri; G.S.Anandha Mala; G.Divya; "Suspicious Data Mining from Chat and Email Data" International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume- 2, Issue-2, April-2014.