ENHANCING THE DIGITAL DATA RETRIEVAL SYSTEM USING NOVEL TECHNIQUES

[1]S.GOWRI, [2]G.S.ANANDHA MALA, [3]G.DIVYA

[1]Research scholar, Department of Information Technology, Sathyabama University, INDIA
[2]Department of Computer Science, St. Joseph's College of Engineering, INDIA
[3]Student, Department of Information Technology, Sathyabama University, INDIA
E-mail: [1]gowriamritha2003@gmail.com, [3]mmcldjs@gmail.com

ABSTRACT

The increase of crime rate around the globe projects the accuracy lag in the current techniques which are followed in process of retrieving and analyzing stored information in the cache gathered from varied channels of communication systems for the digital investigation systems. Taking these accuracy lags as a primary issue the principle objective for the development of this strategically organized system by renewing the existing procedural oriented system and to develop a genuine and a user-friendly framework on crime related information mining system which is mainly objected to extract and inspect appropriate information from cache memory which assimilates emails, chat threads and any text messages for the discovery of the criminal tasks and solve the enigma with the help of the validity concealed within the data. This procedural orientated system is built by merging of the three aspects employed to the data which is given as the input for textual evidencing which are the mails, text messages, chat threads, etc. The three procedural features are, the segregation of the body and header part of the textual corpuses containing all kinds of textual proofs from variety of communication channels which is accomplished using PHP script concept of regular expressions and along with which preprocessing of text to the body portion of the input corpus is done by stemming and tokenization processes, in order to increase the reliability in the text mining process and for the convenience to the forensic departments of investigation in criminology. The final aspect is the search technique implemented in this methodology which is built for the most highly effective and efficient data retrieval process. Even though the followed procedure is an ancient technique in this procedure which is the segregation of body and header in a mail, the primary aspect of this methodology focuses on building an efficient search engine for the purpose of effectiveness in retrieval. For the effective search engine a hybrid algorithm inherited from various other old searching algorithms is used for improvement of the system. This algorithm constitutes of three features mainly; it uses bit-parallelism simulation of the suffix automaton of x^R, it is the alternative form of the Reverse Factor algorithm and if the pattern length is not longer than the memory-word size of the system efficiency in rate of retrieval is high. A procedural methodical would be bought by the utilization of this kind of technique to improvise the existing system initiating the searching system which is highly efficacious; the efficiency of this algorithm for searching is evidenced by its complexity of time and memory and with its values of precision and recall. Integrating all the factors specified would assist in easier reviewing of the text documents.

Keywords: *Textual evidences, tokenization, Stemming, Information retrieval, Preprocessing.*

1. INTRODUCTION

Information exchange could be accomplished in various ways through varied communication channels. Transmission of messages could be done through any of the following ways such as mails transfers, chat messengers, social networks sites, online messaging sites and many more. Usage of communication could be done with any of the motioned services.

Communication based on crime is also done with the help of same service providers but by being an anonymous in the channels which is also a digitized technology. Crime rate increases by increasing in unlawful activity like hijacks, bankruptcy, bomb blasts, kidnapping and many infringements which are deliberated from enormous places were their communication is done through various communication channels. Public chat rooms are one of the main aspects in the proxy communication channels utilized by the instigated victims like cyber predators and pedophiles. In the present technology of the investigation systems manual work is very high, which in turn becomes a cause for the inability of handling large data. The communication rate has been increased to a large extent which becomes a strenuous job for the investigators to handle so taking into consideration of the current scenario such an enormous unstructured data and perform the analyses for wary data in it. In the forensic department the approaches and tools they follow and used to the collected data are of ancient methodologies which are not effective enough over 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 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 also been considered and not the message part, apart from that aspect the other aspect is that the systems 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.

In order to improvise the existing system by considering the above complications faced in the forensic department a new system of approach is been developed. For improvising the effectiveness in textual processing and retrieval of information for monitoring the textual evidences in the forensic department, the usage of the initiated procedure accomplishes increase in the accuracy of maintaining 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 the database created and the body part of the message 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 .txt format, for the further purpose of preprocessing of text in-order to make it more reliable for easier understanding and convenient refining for to inspecting over threads and text document of suspicion. A hybrid algorithm inherited from various old search processes is used for the searching in the user's search engine interface, which fascinates the users in the introduced system. The most fascinating part of this algorithm is that the time complexity factor and the precision of and recall factor of retrieving relevant information where the searching time and the process time are been considered for the complexity value calculation. 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 stemming and tokenization in text mining process is been implemented.

2. RELATED WORKS

A detail study shows that there are many techniques which are been followed in the digital systems. ways of approach. There has been demand and more concentration over the process of clustering of data which is the grouping up process which makes the huge corpus into small clusters which could make the unstructured data into structured data, but does not have any influence to make the process time and relevancy of retrieval of information from those clusters either. The search time of the search engines is not been taken into consideration for data retrieval in enormous dataset size. This initiated system shows the process of data segmentation were the data to be processed becomes less and preprocessing of text which is the stemming and the tokenization makes the retrieval more easier and relevancy of retrieval more effective along with which the hybrid algorithm would accompany for the effectiveness in the process time and search time.

In the study for the stemming techniques there are a many algorithms implemented for various IR systems. There was a short summary which was been given by Frakes & Baeza-Yates (1992) which reports the results of the previous study stating level of stemming use. The impact of performance was comparatively low though the stemming was very beneficial but the stemming was not of most important among the common variants. Therefore there was no evidence of effectiveness in performance of retrieval.

In contrast the study initiated by Krovetz (1993) was shown the increase in the performance of stemming in the IR system. But the collection on which the algorithm was implemented had only short documents, which showed a result which makes sense giving the likelihood of exact match. For large data there have been many more novel techniques with more improvement for the retrieval.

Similarly the study for tokenization had stated that the algorithms which were implemented for tokenization was only for the segregation of text in the documents keeping the punctuators and other symbols as the divider for the text document segmentation. Now the modifications in the algorithm has been justified to decryption for suspicious word mining, cause criminals do not use direct wordings in their text and so the normal analyses for suspicious word mining cannot predict on normal search categorization of text in the document, instead tokenization process is introduces in the system.

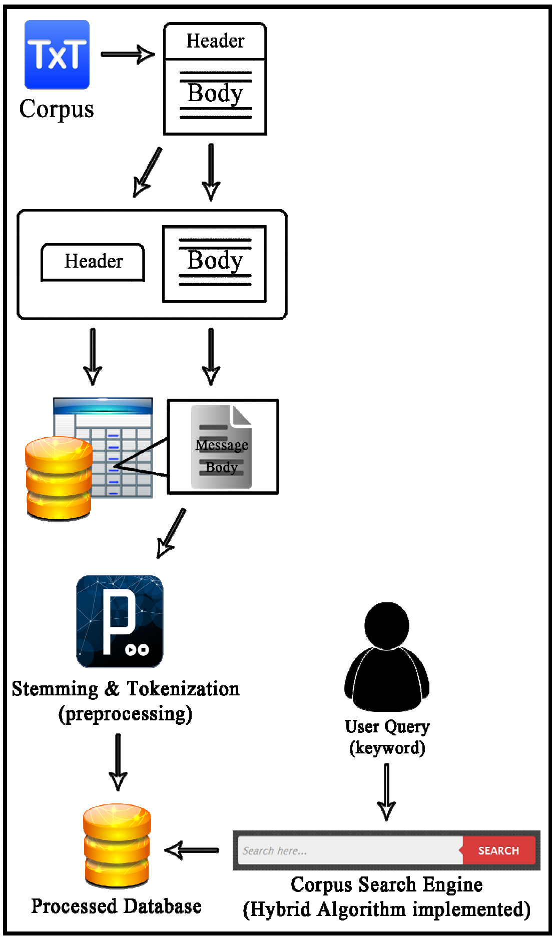


Figure 1: System Architecture

3. SYSTEM ARCHITECTURE OVERVIEW

The overall system overview is set in a user understandable manner which could bring out a user friendly and very effective relevancy retrieval system. First the text corpus is loaded into the created database, this loaded database of mails are then been segmented into header and body in any kind of textually evidenced documents. The segments are then arranged in such a way that the header part is stored directly into the database table with the assigned fields, were the message part is stored as its references in the table, which could be viewed on click of the reference links stored 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. After which the database becomes structured and well organized format, were retrieving from this database is then done by the usage of an hybrid algorithm introduced for the search engine which takes very less process time and search time.

4. EXPERIMENTAL ANALYSIS&RESULTS

4.1. Message

Usually there is only one header is present in any kind of message, which are in the form of fields where a name followed by value is mentioned.

Beginning with a printable character in the header a separate field is set for each line of text. The field name starts with a character usually and when the mail is been extracted to offline dataset and regular expression is set before the field, in the dataset of mail which is take for this analyses had the regular expression of “X-” followed by field name and a separator character “:” which is set to specify the end of field name, where a value is then followed after the end separator. The value is continued onto subsequent lines if spaces or tabs are given as the first character at the value side. 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". It will search for that term in the default field if the field name of a term is not mentioned, that is text, which means "crime" and "text: crime" are equivalent.

4.2. Segmentation of Body and Header from Text Documents

The principle idea for the purpose of 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 and to break the text for preprocessing need, so as to find the accurate retrieval of relevant documents. 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.

Description of PHP and its concept of regular expressions:

a) PHP is an open source language for assembling dynamic web pages.

b) PHP possesses three sets of functions that permit to work with regular expressions.

c) PHP incorporates PCRE by default as of PHP 4.2.0.

d) The most predominant set of regex functions is preg.

e) The former set of regex functions are the ones which start with ereg.

f) PCRE library (Perl-Compatible Regular Expressions) is wrapped around with these regex functions.

g) The functions utilize the POSIX Extended Regular Expressions, which is same as the traditional UNIX egrep command.

h) These functions are predominant for backward affinity with PHP 3.

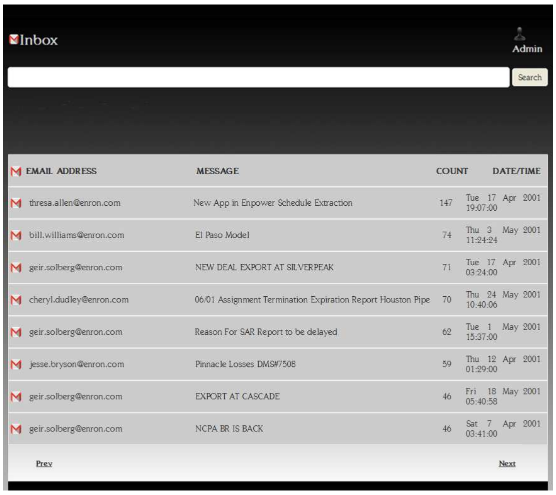
i) The last set is a different from the ereg set, "multibyte" prefix is mb_ to the function names.

j) 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.

k) For regex to handle Far East characters as individual characters, mb_ereg functions are been implemented, or the /u modifier with preg functions.

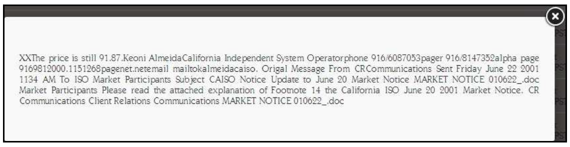
The standard pattern defined for the separation using the regular expressions is:

('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
thresa.allen@enron.com	New App in Enpower Schedule Extraction	147	Tue 17 Apr 2001 19:07:00
bill.williams@enron.com	El Paso Model	74	Thu 3 May 2001 11:24:24
geir.solberg@enron.com	NEW DEAL EXPORT AT SILVERPEAK	71	Tue 17 Apr 2001 03:24:00
cheryl.dudley@enron.com	06/01 Assignment Termination Expiration Report Houston Pipe	70	Thu 24 May 2001 10:40:06
geir.solberg@enron.com	Reason For SAR Report to be delayed	62	Tue 1 May 2001 15:37:00
jesse.bryson@enron.com	Pinnacle Losses DMS#7508	59	Thu 12 Apr 2001 01:29:00
geir.solberg@enron.com	EXPORT AT CASCADE	46	Fri 18 May 2001 05:40:58
geir.solberg@enron.com	NCPA BR IS BACK	46	Sat 7 Apr 2001 03:41:00

Figure 2.1. PHP framework for separated header part.



XXThe price is still 91.87.Keoni AlmeidaCalifornia Independent System Operatorphone 916/6087053pager 916/8147352alpha page 9169812000.1151268pagenet.netemail mailtokalmeidacaiso. Origal Message From CRCommunications Sent Friday June 22 2001 1134 AM To ISO Market Participants Subject CAISO Notice Update to June 20 Market Notice MARKET NOTICE 010622_.doc Market Participants Please read the attached explanation of Footnote 14 the California ISO June 20 2001 Market Notice. CR Communications Client Relations Communications MARKET NOTICE 010622_.doc

Figure 2.2. PHP framework for separated body part.

4.3. Stemming of Segmented Text Document

Stemming of the text document is mainly secured cause of two significant factors, one for making the search of keyword more efficacious and the other for easier comprehending of the document when manually analyzed. Reducing a word to its base is called Stemming (or stem). For example, the words 'processing', 'processed' and 'processes' all have the stem 'process'. A word, or a group of words, and creates the stem, or a group of the stems from the input text.

Stemming is suitable when any kind of text-examination is made. 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 within 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 ordinary text analysis is executed over it, the conclusion of the program would be a text about human, since the only word 'human' apparently occurs more than once in the entire phrase (apart from words like 'i', 'a', etc. which are filtered out usually).

Examining the text after Stemming the text documents, that is replacing all the words 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', which is more than the repetition of the word human.

Algorithm wise the stemming is a problem, because of the presences of enormous rules and notable cases in the English language. But this algorithm for stemming solves the problem of difficulty in utilization due to the introduction of the dictionary factor.

4.3.1. Code process:

Initially a library has to be imported for text based manipulations; here we use JWNL (Java WordNet Library)

Now we can generate a class for the stemming algorithm. The class should contain these members:

Unfortunately, WordNet retrieval 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 low in cost than a constant-time hashmap lookup.

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

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 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: projection
projections
projective → project
projected
projecting

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.

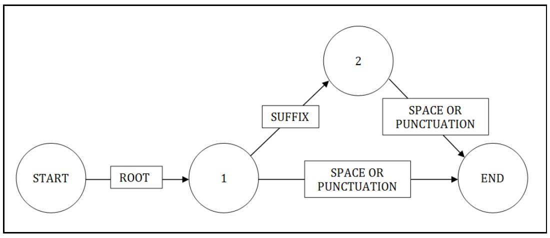
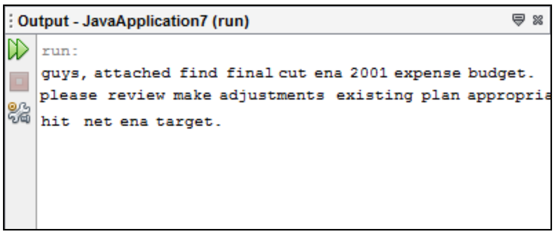


Figure 3: Flow of Stemming Approach



```
run:
guys, attached find final cut ena 2001 expense budget.
please review make adjustments existing plan appropria
hit  net ena target.
```

Figure 4: Stemming Ouput

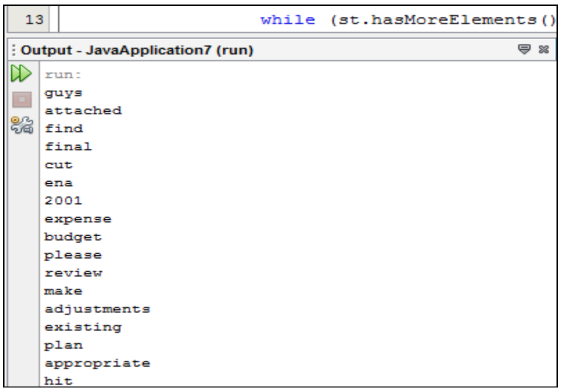
4.4. Tokenization

The process of breaking up of text into lexical units is called Tokenization which is also called as segmentation. The words or number or punctuation mark may be the units. By locating word boundaries in tokenization concept the task is been accomplished. The boundaries of context are meant to be the ending point of the word to the beginning point of the next word. Segmentation is another name.

However, 'word' definition difficulty increases. Simple heuristics is a factor of tokenization:

- All contiguous strings of alphabetic characters are included in tokens, similarly for numbers.
- For the Tokens separation space or line break, or punctuation characters are a few white space characters used.
- The inclusion of punctuation and whitespace may or may not be done in the resultant list of tokens.

```
read(new File("stopwords"))
while(s.charAt(right) != delimiter)
        write(new File(filename))
Output_file= ∑ tokens  //the characters after the
delimiter removal
```



```
13        while (st.hasMoreElements()
```

```
run:
guys
attached
find
final
cut
ena
2001
expense
budget
please
review
make
adjustments
existing
plan
appropriate
hit
```

Figure 4: Tokenization Output

4.5. The Clustering Algorithm for searching technique

In order to filter the mails and display accordingly searching is done by the keyword as an input from the user which is then followed by the sort process to make it clear the content inside the message of the mail. The sort is accomplished using the values of the word_count otherwise called as page ranking which is the count of occurrence of the keyword, 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.

4.5.1. 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 x^R.
c) If the pattern length is not extended than the machine's memory-word size, the algorithm is exceptionally effective.

4.5.2. Description

- This algorithm prompts the BDM algorithm using the bit parallelism.
- ***Properties:***
  - The search is executed in a window which is used to move forward incrementally without looping.
  - When there is no active state the shift of window occurs, which is same as 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:

  $Dj=(Dj\text{-}1 \text{ AND } B[Sj]) << 1$

- Explanation of the update formula
  - No input symbol can re-activate any machine, instead only an active machine can receive a new symbol.
  - The next input symbol regulates on which active machine it will be feasible to try process it
    - The bit mask estimates which machine can execute 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 backward scan of the window reflects the left shit by one.
- Compared to Shift-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

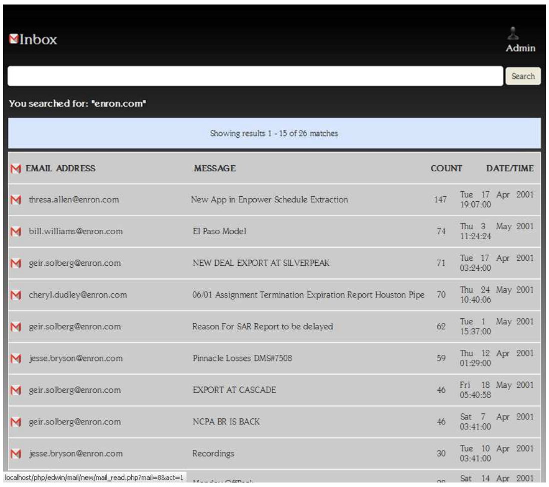
Algorithm: Text search

```
Procedure CA(*x, m, *y, n)
j: {The position of the string present}
s: {Input}
    Begin
    if m > WORD_SIZE
    error("Culturing");
    /* Loading and Initializing */
    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 separate character c, for which a bit mask is saved. 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 retained, 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 imitated a prefix of the pattern in the current window position j, whenever d_{m-1}=1. The shift to the next position is given when the longest prefix is imitated.



You searched for: "enron.com"

Showing results 1 - 15 of 26 matches

EMAIL ADDRESS	MESSAGE	COUNT	DATE/TIME
thresa.allen@enron.com	New App in Enpower Schedule Extraction	147	Tue 17 Apr 2001 19:07:00
bill.williams@enron.com	El Paso Model	74	Thu 3 May 2001 11:24:24
geir.solberg@enron.com	NEW DEAL EXPORT AT SILVERPEAK	71	Tue 17 Apr 2001 03:24:00
cheryl.dudley@enron.com	06/01 Assignment Termination Expiration Report Houston Pipe	70	Thu 24 May 2001 10:40:06
geir.solberg@enron.com	Reason For SAR Report to be delayed	62	Tue 1 May 2001 15:37:00
jesse.bryson@enron.com	Pinnacle Losses DMS#7508	59	Thu 12 Apr 2001 01:29:00
geir.solberg@enron.com	EXPORT AT CASCADE	46	Fri 18 May 2001 05:40:58
geir.solberg@enron.com	NCPA BR IS BACK	46	Sat 7 Apr 2001 03:41:00
jesse.bryson@enron.com	Recordings	30	Tue 10 Apr 2001 03:41:00
			Sat 14 Apr 2001

localhost/php/edwin/mail/new/mail_read.php?mail=8&act=1

Figure 5: Search Result

4.5.3. Precision and Recall

The basic measures used are Precision and Recall for the evaluation of the searching algorithm. There are 'N' numbers of records stored in the corpus which are associated to the search topic. The Records are assumed 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 hybrid algorithm to that of the reverse factoring algorithm which is kind of an alternative form shows the data retrieval which is comparatively higher in 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, 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, which is expressed in terms of percentage.

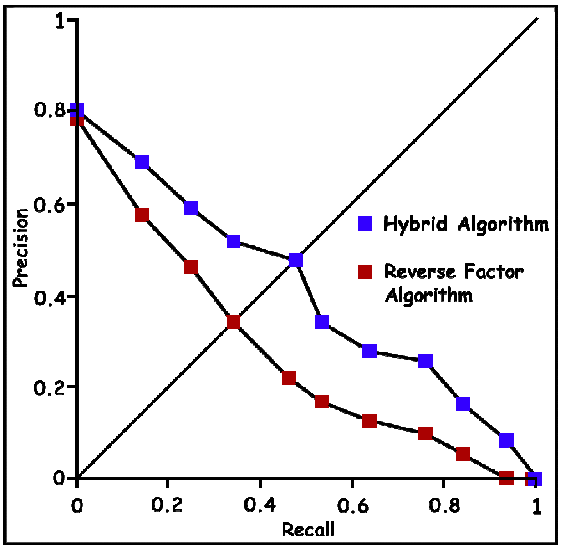


Figure 6: Precision and Recall graph for 50 Quires

5. COMPARISON AND SEARCHING TECHNIQUE

The Fig 7. graph shows the time complexity of different searching techniques, where the vertical axis shows the average time in milliseconds every algorithm needs to preprocess and parse the stemmed and tokenized text document. Thus the algorithm with the least value of time is better. The suggested algorithm has the search time of approximately 11.2 milliseconds and the preprocessing time of approximately 2 milliseconds which would give a rapid performance with an efficacious retrieval of relevant documents, as mentioned with the precision recall graph for the the algorithm.

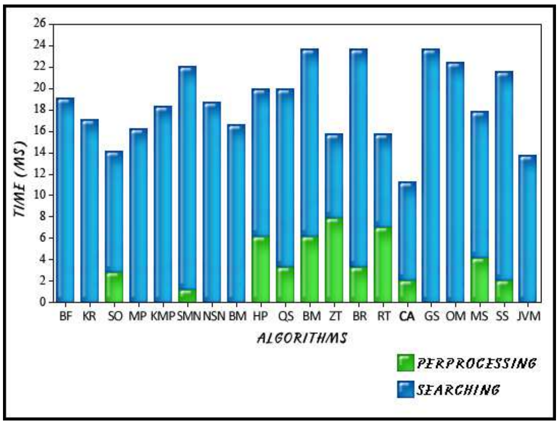


Figure 7: Comparison chart for Searching Algorithms

6. LIMITATIONS OF PROPOSED SYSTEM

- In the algorithm for the search engine the pattern length should not be extended than the machine's memory-word size.
- If the stemming is done after tokenization there would be a loss of data in the processed document, therefore the preprocessing is set in such a way that the stemming is only done first then tokenization.

7. CONCLUSION

In this paper the performance of analyzing and mining (stemming and tokenization) over the collected data on textual documents like mails, message threads, chat threads etc., for which a system has been produced in which the segmentation of the header and body part of the corpuses is done to the dataset and then the segmented body part is then manipulated using stemming and tokenization preprocessing techniques , which are implemented in java for a better viewing and searching process of keywords and phrases, the searching process used in this system is the hybrid algorithm derived from various other ancient techniques and implement process. Integrating these entire factors into a single unit of system initiating the use of performance to find out the information related to crime in the forensic investigations in digital text retrieval. As per the work progress the test over mail corpus is 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 by introducing the concept of document clustering concept which is been accomplished using Google's crawling concept is to be initialed in later further future enhancement which needs to be worked on which will enable to perform auto detection of suspected text documents on the basis of training set and samples provided to it even an efficient inter cluster sorting technique and mapping of clusters is also going to be introduced in the future enhancement process.

REFRENCES:

[1] 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

[2] S.Sathya Bama, M.S.Irfan Ahmed, A.Saravanan; "A Mathematical Approach for Improving the Performance of the Search Engine through Web Content Mining"; Journal of Theoretical and Applied Information Technology 20th February 2014. Vol. 60 No.2

[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] Dr. Khalaf Khatatneh, Iman Hussein "Information Retrievals Tries Tree Vs Inverted File Word Method for Arabic Language"; Journal of Theoretical and Applied Information Technology 2010

[5] 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).

[6] 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.

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

[8] 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.

[9] 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).

[10] Navarro G., Raffinot M., "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, 14-31-1998.

[11] 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.

[12] David A. Hull "Stemming Algorithms- A Case Study for Detailed Evaluation" June 7th 1995

[13] 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.

[12] 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.

[14] Igor Santos, Carlos Laorden, Borja Sanz, Pablo G. Bringas; "JURD: Joiner of Un-Readable Documents algorithm for Reversing the Effects of Tokenization Attacks against Content-based Spam Filters"