Journal of Engg. Research **Vol. 3 No. (4) December 2015** pp. 43-58

تحسين الإطار الاستراتيجي التعاوني للحداثة في استرجاع رقمي للبيانات النصية عن طريق النشر والخوارزميات في مجال التحليل في الطب الشرعي

***س. جواري و**ج. س. أناندا مالا**
*قسم تقنية المعلومات، جامعة ساثيباما، تشيناي، الهند
**قسم علوم وهندسة الحاسب الآلي، كلية أساري للهندسة، تشيناي، الهند

الخلاصة

يقدم هذ البحث لخوارزميتان متقدمتان ومدمجتان في نظام متكامل. الأول هو خوارزمية المسار الديناميكي الاختياري لتجميع الوثائق والآخرى هي نافذة ثنائية مؤجلة لمحرك البحث الشخصي، وهي خوارزمية مسار ديناميكي والتي تم اشتقاقها من مفهوم تقنية حفارات الجوجل المطبق حاليا في معالجة البيانات خارج خدمة الشبكات والخوارزمية نافذة ثنائية لمحرك البحث مشتق من تقنيات خوارزميات البحث الأخرى. وجرى إنجاز النظام المقترح لإعطاء بنية بيانات مناسبة لمحتوى البيانات الداخلة، والبيانات المستخدمة كمدخل هو مجموعة البيانات انرون والتي هي كبيرة الحجم وغيرمنظمة. وتم تصميم النظام بمساعدة طريقة دمج جميع الوحدات الفردية والمستقلة في النظام من خلال جمعهم تحت إطار واحد. والهدف من الوحدات هي تجهيز البيانات، وتوثيق المجموعات ورسم خرائط التجمعات ومحرك البحث. هذا النظام مع تكرير الإطار الدقيق المتكامل ومن المرجح أنها الطريقة الأفضل، لأن تعريف بسيط لنظام استرجاع البيانات يؤثر على اتساق استرجاع المعلومات الغير ذات صلة. رغم أن هناك الكثير من النظم القائمة في قسم الطب الشرعي مع التعريف البسيط لمحركات البحث، وينظر في الاسترجاع إلى حد كبير إلى عدم الاعتداد من دون أي عمليات أخرى. وبالتالي فإن تصميم هذا النظام المتكامل والتي هي عملية آلية باستخدام وحدات تكوين معرفة المعالم والمذكورة أعلاه. وهذا النهج المنظم للاستخدام كافة الأدلة النصية الرقمية، والتي تساعد في سرعة تحديد الجريمة. ويتم تحليل نتائج النظام المقترح عن طريق الحصول على قيم دقيقة ومقارنتها مع نتائج محركات البحث لاختبار فعالية في معدل استرجاع البيانات.

Strategic enhancement of the collaborative framework for novelty in retrieval from digital textual data corpus by deploying DPSC and RBWM algorithms for forensic analysis

Gowri Shanmugam* and Anandha Mala Ganapathy Sankar**

**Department of Information Technology, Sathyabama University, Chennai-600119*

***Department of Computer Science and Engineering, Easwari Engineering College, Chennai-600089*

**Corresponding author: gowriamritha2003@gmail.com*

ABSTRACT

This paper proposes two advanced algorithms embedded into an integrated system; one is a Dynamic Path Selection Clustering (DPSC) algorithm for the document clustering and the other is the Rearward Binary Window Match (RBWM) algorithm for the user's search engine. The DPSC algorithm is derived from the concept of Google's crawler technique implemented in offline processing and the RBWM algorithm for search engine is derived by utilizing the techniques of other search algorithms. The proposed system is being accomplished for giving an appropriate data structure to the input dataset content. The dataset used as input is the Enron dataset, which is large in volume and unstructured. The system is designed with the help of integrating all the individual and independent units into a system by bringing them under one frame and the units are data preprocessing, document clustering, mapping of clusters and search engine. This system, with fine refining integrated frame, would likely evidence in a better way, since simple definition of the system for data retrieval affects the consistency of irrelevant information retrieval for evidencing to be increased. Though there are plenty of existing systems in forensic department with only simple definition of search engines, without any other processes the irrelevancy in retrieval is seen to a larger extent. Consequently, a design of this integrated system, which is automated in process by using the above well defined configured units, is proposed. This systematic approach is for adequate use of digital textual evidences, which assists in quicker crime identification rate. The outcomes of the proposed system are analyzed by obtaining the precision and recall values and comparing them with the results of Metasearch engines like Dogpile and Metacrawler, to test the efficacy in retrieval rate.

Keywords: Data management; document clustering; Google's Crawler; preprocessing; semantic.

INTRODUCTION

The process of allowing transformation of large volume of heterogeneous and raw data into genuine information is called data morphing, whereas data mining is the examining step of the "Knowledge Discovery in Databases (KDD)" procedure. It is an interdisciplinary subfield of computer science and is the computational procedure of discovering patterns in huge datasets involving techniques at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extricate information from datasets and transform it into a semantic structure for further use. Besides the raw analysis step, it involves database and data management characteristic, data preprocessing, model and inference deliberations, interestingness metrics, complexity considerations, post-processing of structures discovered, visualization, and updating online. Data mining tasks are basically categorized into two major tasks, one of which is the descriptive task, whose purpose is pattern derivation (correlations, trends, clusters, trajectories, and anomalies) and another is the predictive task, where its resultant is obtained by comparing one attribute value to the other. The methodology of organizing a larger text dataset based on a category is done in such a way that the objects in one cluster are not similar to that of the objects in other clusters. Apart from the specified tasks, there is a fundamental task in the field of text mining, which is usually concerned with relating the similarity between the data grouped into clusters.

Digital textual analysis is one of the most significant activities to be implemented in the forensic department, where the increase of textual evidences increases through various network communication channels (ÁlvaroCuesta *et al.*, 2014) like SMS, mails, public chat rooms, etc. Acquisition of precise data and fast retrieval would favor the applications for evidencing information from large data corpuses of digital textual data. From the analysis of several criminal acts in the present days, it would predict that the communication between people (criminals of ordinary people) mostly takes place through the unsecured digital text channels using anonymous names. Such communications are evidences for most of the cases, where processing of such data could be done in our proposed framework. Therefore, as a part of future work, automation in monitoring continuously over network communication channels is being carried out. Our proposal in this paper focuses on the automation of workload in preprocessing and structuring unstructured data.

OBJECTIVE

The main aim of this system is to minimize the time taken for the data retrieval on search hit and to increase the rate of relevancy of retrieved data. These aspects are accomplished with the help of this system, specifying the new methodology of clustering technique using DPSC algorithm, which is structurally efficacious. Further, mapping of clusters is done to give a semantic relationship and finally a hybrid search algorithm for the search engine interface is deployed. This application aims at giving a well defined structure to the input corpus by processing and filtering all files. For this system, a frame is set without the requirement of any prerequisites other than Java platform (being an offline desktop application).

RELATED WORK

Over many years, network traffic for digital textual data transfer like texting, e-mails, chat threads and many other modes between nodes has become enormously large. This enlarging of digital textual data makes the retrieval of information difficult for analysis. Thus, many approaches were developed, either by various clustering techniques or by the use of other methodologies for handling the data. It is generally difficult to compare different approaches, as they vary in feature selection, choice of supervised or unsupervised classification algorithms, and set of classified traffic classes.

The analysis of few search engines is shown in the comprehensive survey by Sampath & Pavithra (2010). One complicated comparison between different studies is the fact that, classification of performance depends on how relevant the retrieved data is from the test data used, to evaluate the accuracy in relevancy. The factors used for the purpose of comparison are the precision and recall values. Viewing into the search engines of forensic department, Nicole & Jan (2007) stated that, the current digital forensic text string search tools use match and/or indexing algorithms to search digital evidence at the physical level to locate specific text strings and are designed to achieve 100% query recall. Although giving the nature of the dataset, extremely high incidences of hits that are not relevant to investigative objectives have been obtained. Though internet search engines suffer in a similar manner, they employ ranking algorithms from the user's perspective to present the search results in a more effective and efficient manner. Current digital forensic text string search tools failed to group or order search hits in a manner that improves the investigator's ability to get to the relevant hits first. They had proposed and empirically tested the feasibility and utility of post-retrieval clustering of digital forensic text string search results, by using Kohonen Self-Organizing Maps; a self-organizing neural network approach. There are plenty of encryption protocols; for instance the methodology proposed by Nam-Su & Dowon Nam-Su (2013). It has a new searchable encryption protocol with a conjunctive keyword search, based on a linked tree structure instead of public-key based techniques.

There are several approaches formed by the concept of clustering big data. Most of the clustering techniques have been derived from old techniques, which are the k-means (Suiang-Shyan & Ja-Chen, 2012), fuzzy (Tasić & Stojanović, 2006; Rathna & Sivasubramanian, 2014; Sendilkumar *et al.*, 2013), hierarchical, density based clustering, distribution based clustering, centroid based clustering, etc. The approach of flow cluster is unique from all other utilized concepts, wherein the clusters of the traffic is formed over packet transmission channel using a parameter as a base, as proposed by Anthony *et al.* (2004). The usage of a variety of features such as packet length statistics, inter arrival times, byte counts and connection duration for the purpose of clustering was a unique perspective on this approach. Expectation-maximization (EM) clustering (Sebastian *et al.*, 2005) was used for the purpose of grouping together the flows, which would have similar features. The computation of the reduced feature sets (a few have been deleted) helps to simplify classification for the cluster, which represent the set of traffic classes. The concept utilized for clustering is the distance between the nodes of the femtocell network proposed by Hong & Rongfang (2014). Wei Kuang *et al.* (2014) developed a novel cluster-based routing protocol for corona-structured wireless sensor networks. Based on the relaying traffic of each cluster head conveys, adequate radius for each corona can be determined through nearly balanced energy depletion analysis, which leads to balanced energy consumption among cluster heads. Simulation results demonstrate their clustering approach effectively improves the network lifetime, residual energy and reduces the cluster head rotations in comparison with the MLCRA protocols.

A well known factor in big data of the current scenario is the concept of semantics, which plays a vital role over the novel approaches. Seung Ryul & Imran (2014) have covered a literature survey on semantic-based approaches and tools, which can be leveraged to enrich and enhance today's big data. They presented an elaborate literature on 61 studies during the period 2011 to 2014, by highlighting the key challenges that the semantic approach needs to address in future. Now taking these instances into consideration, they presented cutting-edge approaches to ontology engineering, ontology evolution, searching and filtering relevant information, extracting and reasoning, distributed (web-scale) reasoning, and representing big data. This system is incorporated with the approach to ontology for the clusters formed. Consequently the examination of various works related to data mining concepts had assisted in designing the proposed framework.

OVERVIEW OF PROPOSED SYSTEM ARCHITECTURE

From the analysis of above mentioned recent works, the significance of IR framework and techniques embedded in them were studied and consequently architecture was designed for the proposed Framework. The proposed system architecture is shown in Figure 1 below and the description of each unit is given in the subsections that follow.

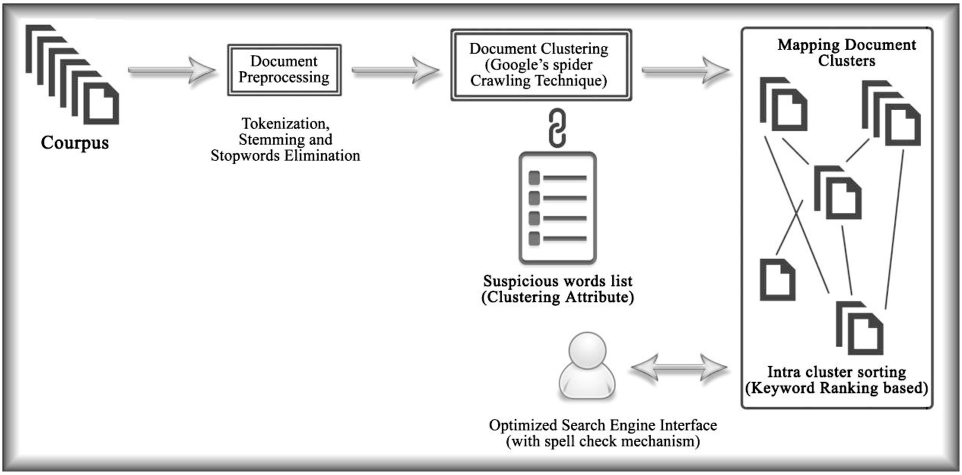
Courpus
Document Preprocessing
Tokenization, Stemming and Stopwords Elimination
Document Clustering (Google's spider Crawling Technique)
Suspicious words list (Clustering Attribute)
Mapping Document Clusters
Intra cluster sorting (Keyword Ranking based)
Optimized Search Engine Interface (with spell check mechanism)

Fig. 1. Architecture diagram of proposed system

Mail Corpus Loading and Segmenting Unit

The mails are initially loaded by choosing the dataset from the location of storage, after which segmentation is done. The principle idea of segmentation of the header and the body of a text document is to give a clear analysis over both the address, the message information and to break the text for preprocessing, so as to find accurate retrieval of relevant documents. The head part of the mail is stored in the table with the respective columns and the body part is stored in ".txt" format, with its reference in the table. The dataset used as input to the system is taken from the Enron corpus. Only one header is present in any kind of message, where a name followed by value is mentioned in each field. A separate field is set for each line of text, beginning with a printable character in the header. The field name very often starts with a character and a regular expression is set before the field on extraction to offline dataset. The regular pattern of expression for this dataset is "X-" followed by field name and a separator character ":", which specifies the end of the field name, after which the value is followed. If spaces or tabs are given as the first character at the value side, the value is continued onto subsequent lines. For the field's name and values 7-bit ASCII characters are only allowed, where the MIME encoded words are used for representing the non ASCII values.

The actual storage pattern of all the fields of any mail has been prefixed with the "x-" as a separator variable, which helps to segregate the mail's header and body part easily during the segmentation and storage in database. Field names are case sensitive. For example "Subject" and "subject" will not be treated as same.

The standard pattern defined for the separations are: "x-from:", "x-to:", "x-cc:", "x-bcc:", "x-folder:", "x-origin:", "x-subject:", "x-filename:" "x-timestamp:"

The time stamp of the emails which are sent, received or forwarded is of the form (YYYYMMDDhhmmss).

Preprocessing unit

The purpose of preprocessing unit is that the time taken for any processing of text document to become faster. This is because after preprocessing the content of the document formed from the original text document will reduce by elimination of unwanted or very commonly used contexts. Those words are only parsed, whenever any process takes place. As a result, the relevancy also becomes high, as parsing over required data is done and only relevant documents are retrieved, when search hit is accomplished.

The three linguistic morphologies used for the pre-processing are: tokenizing, stemming and stopwords elimination (Gowri *et al.*, 2014a), which are explained as follows. (i)Tokenization: Breaking of text into units of characters and words. (ii) Stemming: The process of reducing inflected (or sometimes derived) words to their word stem, base or root form - generally a written word form. (iii)Stopwords Elimination: This is the final process of preprocessing units. The unwanted words filtered out are the stopwords. No explicit index list of stopwords is present in common, were such filters are not always used. To support phrase search, some tools particularly avoid removal.

Dynamic Path-Selection Clustering algorithm

Most of the experimental evidences prove that IR applications are largely benefited by the clustering process, in order to overcome the problems, which arise in search engines in order to benefit the user by providing relevant information to be retrieved and to improve navigation and retrieval performance. Usually clustering is done by following factors; mainly similarity, distance between nodes, marginal coverage or centroid factor of the datasets.

In the current document clustering algorithms, the representation of the documents is done by the usage of vector space model (VSM) (Subhashini & Senthil, 2011) alone, which represent documents as vectors in the space of terms and uses the cosine similarity between the document vectors to estimate their similarity. Semantic relationships between the terms have been ignored. In this case, relevant documents cannot be retrieved by just matching the terms in the query to the document. Further, the clustering algorithms were more enhanced, based on cluster similarity of nearest neighbors, so that it will not only retrieve the documents which contain the query terms, but also retrieve documents which are similar to the retrieved ones. This process id is done by finding the ‘n’ nearest neighbors of all points with a certain similarity threshold.

This irrelevancy had motivated in bringing up a new proposal over the enhancement in document clustering technique by the usage of the most popular Google's spider crawling technique (Rudi & Paul, 2007) introduced into the clustering process for an adequate search process. The documents are then ranked, based on the term frequency. The system even includes an offline thesaurus application adaptor. This thesaurus is utilized to check for the synonyms of the user given query for relevancy in retrieval from the dataset, as well as for grouping by semantic correlation to be done. This improves the accuracy of relevant retrieved documents.

Google basically focuses on the link structure of the websites to determine and index all the web links. Link structures are the processes set for visiting all the web pages with the help of links in the visited web pages. In the proposed algorithm, the link structure of the document formed after the preprocessing unit is given, by considering one of the following criteria:

1) The mailer ids sent, forwarded, CCed, BCCed or received.
2) The relation between the unique ids given to each mail.
3) The word of suspicion used in each mail.
4) The timeline property of mails in the threads.

Figure 2 below explains the working of the DPSC Algorithm for clustering of mail corpus.

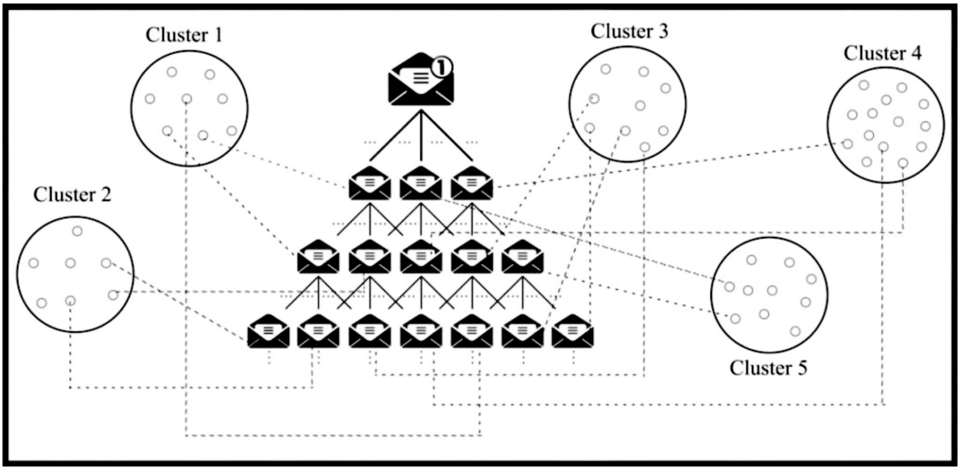

Fig. 2. Clusters formation for mails

Algorithm:

The following sequences of events are followed for clustering of documents:

1) A pointer is assigned for parsing through the textually manipulated files.

2) The parser is simultaneously checked with the list of suspicious words created.

3) Word count is also done as a simultaneous process by the parser.

4) Path redirection of files is done as the parser fetches the suspicious words in the text.

5) A cluster head becomes the node with the highest weight of the word count.

The mathematical representation of the working of proposed system is given below:

$$N = \sum_{i=0}^{\mathrm{EOF}} (A[i] + L) \begin{cases} A[i \text{ to } i+l] = \mathrm{word}[j \text{ to } j+l] \rightarrow \text{do index document to folder} \\ A[i \text{ to } i+l] = \mathrm{word}[j \text{ to } j+l] \rightarrow i++ \text{ (positional icremental)} \\ \qquad\qquad A[i] = \mathrm{EOF} \rightarrow \text{Exit} \end{cases}$$

$$R[n] = \log_{10} N$$

Where

‘R[]’ is the log of the resultant list of clustered sets

‘word[]’ is the word form suspicious word list

‘N’ is the summed up the value

‘n’ is the size of the cluster

‘i’, ‘j’ is the index variable which is used for parsing of the file

‘A’ is the start position of the word and ‘L’ is the size of (Word)

‘EOF’ represents the End Of File

The word length is measured by the separator token ‘space’ before the word and after the word.

Semantic mapping of clusters

The mapping of clusters is done by creating an ontology, which gives a structural framework by organizing the clusters. Ontology is defined as a formal, explicit specification of a shared conceptualization, which provides a common vocabulary to denote the types, properties and interrelationships of concept in clusters. Figure 3shows an example of ontology map.

The tool used for generating the mapping is the protégé 5.0 tool. The purpose of mapping the clusters is to bring in a structure to the clusters formed. This would make the fetching of data more efficacious. The relational factor used to map the clusters is the keywords and their weights, which are listed in the document formed for all the mails.

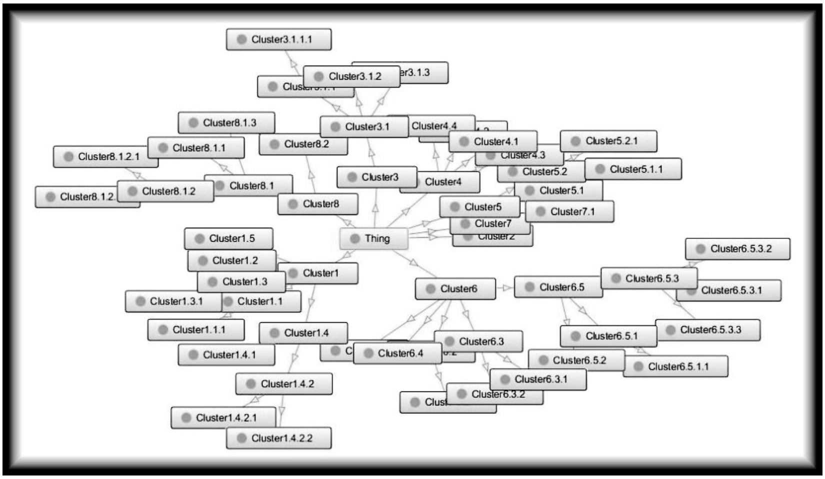

Fig. 3. Sample ontology for clusters mapping

Rearward Binary Window Match

There are plenty of problems, when the present searching techniques are chosen and the following are a few problems, which have been examined. Based on the precision and recall values of the search, it is found that the precision value of the search engines is considerably low, which means that the results of the search engine are irrelevant, which in turn increases the difficulty to find the relevant information. The other problem is that the recall value is also very low, leading to difficulty in indexing all the appropriate information available over the dataset, which might also lead to loss of relevant indexing information. (Gowri *et al.*, 2014b).

Significant features

a) This algorithm is the combination of the Reverse Factor Algorithm (RF) and Backward Non-Deterministic Dawg Matching Algorithm (BNDM).

b) Bit-parallelism simulation is employed for suffix automaton of x^R and conversation of string to bits has been chosen from the concept of reverse factoring algorithm.

c) If the criterion of pattern length not extending the machine's memory-word size is satisfied, the algorithm is exceptionally effective. As a result dynamic variation in the size of the window is done, which has been altered with the help of the word size.

Description

- The search is executed in a window used for incremental forward moment without looping.
- When there is no active state (that is the flag value is altered after permutation

trails), the shift of window occurs, which is same as in the concept of BDN, since no looping process is done.

- The Search status updating formula is: $D_j = \left(D_{j-1} AND\ B[S_j]\right) \ll 1$
  - No input symbol can re-activate any machine, instead an active machine receives new symbol.
  - The next input symbol regulates on which active machine will be feasible to try to process it.
- The bit mask estimates which machine can execute the input symbol (those having the corresponding 1 in the bit mask, i.e. b_i=1)
- To remain active, a machine must be previously active (d_i=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 shift by one.
- Compared to Shift-And and Shift-Or the bit mask has reverse order, since backward search.

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

Algorithm:

1) A window with a dynamic memory allocation is set for the purpose of search (the dynamic allocation is done because the size of the window depends on the word for comparison)
2) The word search is made without looping process; instead the window is moved, when the character wise comparison does not match the hit word.
3) When the word is found in the clusters, the count value of the searched word is also set to increment and the result is displayed along with the word count in a sorted manner.

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 and 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 of the key word which is to hit. 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.

EXPERIMENTAL RESULTS AND DISCUSSIONS

The two search engines acquired for the purpose of analysis are Metasearch engines (Ya-li *et al.*, 2010). Unlike all other search engines, they are built by integration of many other search engines like Google, Yahoo search, Ask, MSN search, About, MIVA, Looksmart and more. In these search engines, searches can take place simultaneously. Furthermore, they retrieve the results from each search engine and combine into a single list, avoiding the factor of redundancy. These search engines are stated to be effective in the analysis for the purpose of comparison to existing ones, because of the utilization of almost all web search engines. Consequently the property of diversity shows that the Metasearch engines are effective ones. Therefore, these Metasearch engines are considered for our experimental analysis, as they have the property to search in specific path (set to an Enron dataset for experiment).

In this study, only text search with 15 random queries related to Enron data was taken as input query from all the 3 search engines, as the proposed search engine is set only for text document retrieval. The other criteria of the search engine are also set only to English language, because search engine for other languages is in progress as future work. As for real time analysis, since there is enormous number of mails, the consideration of an Enron dataset is used as an input of data, from about 150 users in the framework.

Comparison of precision values acquired on evaluated analysis

The precision value is basically the measure of the number of relevant documents retrieved by a search engine on a search hit. The proposed clustering technique uses Google's Crawler method for formation of clusters in this case; but the original crawler is used for indexing of sites, which have been spotted on the sites the spider parses. However, in our proposal the crawler has been used for fetching of words in the documents formed from mails of text and according to fetched words, the file is placed in its appropriate folder, which is dynamically set.

Table 1 shows the precision values of Proposed Search, Metacrawler, and Dogpile, as the analysis show the mean value of all the three search engines are 0.71, 0.59 and 0.63 respectively. This shows that the resultant or the proposed system is comparatively better than the other 2 search engines. Figure 4 shows a graphical representation of tabulated precision values and the comparison of the maximum peak values (0.88, 0.86 and 0.81 respectively). Also, it explains that the proposed system is comparatively better than Metasearch engines.

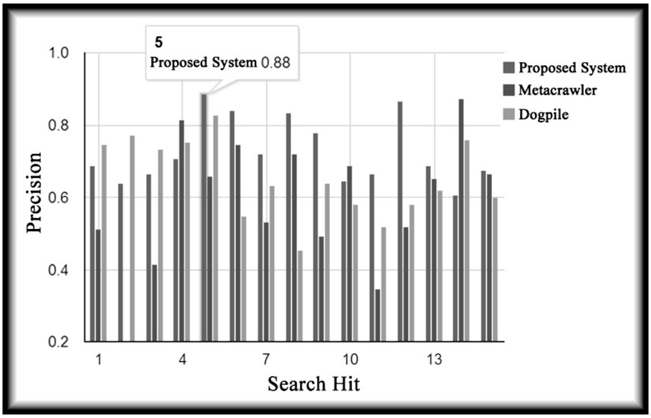

Fig. 4. Precision of Proposed Search, Metacrawler and Dogpile

Table 1. Precision values of Proposed Search, Metacrawler and Dogpile

Search Queries	Proposed Search	Metacrawler	Dogpile
Q#1	0.66	0.48	0.73
Q#2	0.61	0.37	0.76
Q#3	0.64	0.8	0.71
Q#4	0.69	0.64	0.74
Q#5	0.88	0.73	0.81
Q#6	0.83	0.5	0.51
Q#7	0.7	0.7	0.61
Q#8	0.82	0.46	0.41
Q#9	0.76	0.66	0.61
Q#10	0.62	0.3	0.55
Q#11	0.64	0.49	0.49
Q#12	0.86	0.63	0.55
Q#13	0.66	0.86	0.59
Q#14	0.58	0.64	0.74
Q#15	0.65	0.62	0.57
Average	**0.71**	**0.59**	**0.63**

Comparison of recall values acquired on evaluated analysis

Relative recall is the measure of number of relevant documents retrieved from the total number of relevant documents in the database. Thus the ability of retrieval among the three search engines has been analyzed. Table 2represents the recall values of the 3 search engines for 15 search hits. It is evident from the above values of retrieval that the overall mean of the recall values is high for the proposed system compared to results of other 2 Metasearch engines.

The recall values are 0.71, 0.50 and 0.49, which show that the search engine retrieves at most of 71%, 50% and 49% of the relevant data on search hits respectively. The peak of the maximum values of the 3 search engines are 0.98, 0.58 and 0.60 respectively (as shown in the Figure 5), where this comparison also shows that the proposed system has a better retrieval rate over the other search engines.

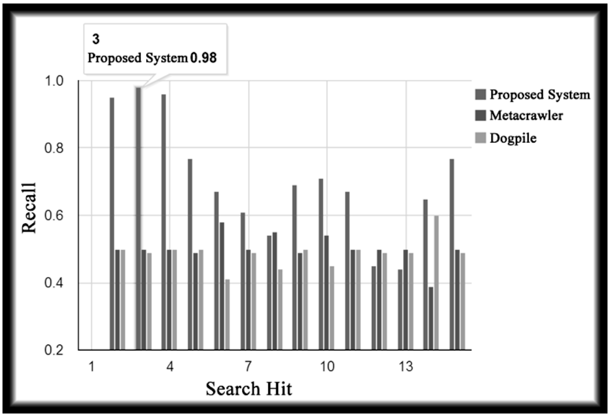

Fig. 5. Relative Recall of Proposed Search, Metacrawler and Dogpile

Table 2. Relative Recall of Proposed Search, Metacrawler and Dogpile

Search Queries	Proposed Search	Metacrawler	Dogpile
Q#1	0.82	0.50	0.50
Q#2	0.95	0.50	0.50
Q#3	0.98	0.50	0.49
Q#4	0.96	0.50	0.50
Q#5	0.77	0.49	0.50
Q#6	0.67	0.58	0.41
Q#7	0.61	0.50	0.49
Q#8	0.54	0.55	0.44
Q#9	0.69	0.49	0.50
Q#10	0.71	0.54	0.45
Q#11	0.67	0.50	0.50
Q#12	0.45	0.50	0.49
Q#13	0.44	0.50	0.49
Q#14	0.65	0.39	0.60
Q#15	0.77	0.50	0.49
Average	**0. 71**	**0.50**	**0.49**

CONCLUSION

Today's search engines are most effective search tools for millions of users around the world. Numerous retrievals are done from each search engine every day, which are either relevant or irrelevant data. Especially, retrieval at the forensic department is also relatively high, as conversations of intruders are also being done through the network channel itself. Therefore, this new system is initiated to benefit forensic department, with textual evidences collected and stored in the corpus, which have been processed and retrieved in an effective and faster manner. An evaluation test with the factors of precision and recall was done with the designed system, which has resulted as demonstrated, shows that the precision and recall values of the proposed system are comparatively higher than those of the Metasearch engines.

FUTURE ENHANCEMENT

For future enhancement, the system is made to work with following processes, which have a greater effect over the resultant, than the proposed work. The process is multilingual search, enabling the system to search over different languages in addition to English. Also, automation in collection of data from the network channel directly eliminates manual work entirely. For reducing complications on the user's side, following 3 processes are to be enabled.

- Word or phrase prediction during search, so that there would be no need of typing the entire phrase, instead could rather choose suggested alternative phrase or word, when result is not found. This phase could help the users to go for alternatives and alert them, if content is not found.
- Preservation of history of number of times viewed could help the user to easily view the document again, in addition to giving an explicit search.
- In the language process, Lemmatization will be introduced for morphing which would return meaningful sentences on removal of stop words.

This integration would help in improvement of the overall process system.

ACKNOWLEDGMENTS

We thank Sathyabama University for providing us with various resources and unconditional support for carrying out this work.

REFERENCES

Anthony McGregor, Mark Hall, Perry Lorier and James Brunskill 2004. Flow clustering using machine learning techniques. Proc. of 5th Int. Workshop on Passive and Active Network Measurement.

Sebastian Zander, Thuy Nguyen and Grenville Armitage 2005. Self-learning IP traffic classification based on statistical flow characteristics. Proc. of 6th Int. Workshop on Passive and Active Measurement.

Tasić D. S. and Stojanović M. S. 2006. Modified Fuzzy Clustering Method for Energy Loss Calculations in Low Voltage Distribution Networks. ELEKTRONIKA IR ELEKTROTECHNIKA. **2**(66): 50-55.

Nicole Lang Beebe and Jan Guynes Clark 2007. Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results. The International Journal of Digital Forensics & Incident Response. **4**: 49-54.

Rudi L. Cilibrasi and Paul M. B. Vitányi 2007. The Google Similarity Distance. IEEE Transactions on knowledge and Data Engineering, **19**(3).

Sampath Kumar B. T. and Pavithra S. M. 2010. Evaluating the searching capabilities of search engines and metasearch engines: a comparative study. Annals of Library and Information Studies. **570**: 87-97.

Ya-li Cao, Tie-jun Huang and Yong-hong Tian 2010. A ranking SVM based fusion model for cross-media meta-search engine. Journal of Zhejiang University SCIENCE C. **11**(11): 903-910.

Subhashini, R. and Senthil Kumar, V. J. 2011. A framework for efficient information retrieval using NLP techniques. Communications in Computer and Information Science. **142**: 391-393.

Suiang-Shyan Lee and Ja-Chen Lin 2012. An accelerated K-means clustering algorithm using selection and erasure rules. Journal of Zhejiang University SCIENCE C. **13**(10): 761-768.

Nam-Su Jho and Dowon Hong 2013. Symmetric Searchable Encryption with Efficient Conjunctive Keyword Search. KSII Transactions on Internet and Information Systems (TIIS). **7**(5): 1328 - 1342.

Sendilkumar S., Mathur B. L. and Mohammed Imran 2013. Discrimination of Power Transformation inrush and internal Fault Current using Time to Time Transformation and Fault Classification using Fuzzy Clustering. Journal of Engg. Research. **1**(3): 87-108.

Álvaro Cuesta, David F. Barrero and María D. R-Moreno 2014. A Framework for Massive Twitter Data Extraction and Analysis. Malaysian Journal of Computer Science. **27**(1): 50-67.

Gowri S, Anandha Mala G.S and Divya.G 2014a. Text Preprocessing for the improvement of Information Retrieval in Digital Textual Analysis. International Conference on Mathematical Science(ICMS 2014) Sathyabama University- Elsevier, pp. 174-179.

Gowri S, Anandha Mala G.S and Divya.G 2014b. Enhancing the Digital Data Retrieval System Using Novel Techniques. Journal of Theoretical and Applied Information Technology. **66**(2).

Hong Wang and Rongfang Song 2014. Clustering Based Adaptive Power Control for Interference Mitigation in Two-Tier Femtocell Networks. KSII Transactions on Internet and Information Systems (TIIS). **8**(4): 1424-1441.

Rathna R. and Sivasubramanian A. 2014. Energy Conservation in Radiation Monitoring. Journal of Engg. Research. **2**(2): 123-138

Seung Ryul Jeong and Imran Ghani 2014. Semantic Computing for Big Data: Approaches, Tools, and Emerging Directions (2011-2014). KSII Transactions on Internet and Information Systems (TIIS).**8**(6): 2022 - 2042.

Wei Kuang Lai, Chung-Shuo Fan and Chin-Shiuh Shieh 2014. Efficient Cluster Radius and Transmission Ranges in Corona-based Wireless Sensor Networks. KSII Transactions on Internet and Information Systems (TIIS). **8**(4): 1237-1255.

Enron dataset- http://www.cs.cmu.edu/~enron/

Crawling & Indexing-http://www.google.com/intl/en/insidesearch/howsearchworks/crawling-indexing.html

Google Crawlers- https://support.google.com/webmasters/answer/1061943?hl=en

Web crawler- http://en.wikipedia.org/wiki/Web_crawler

Submitted: 21-10-2014

Revised: 16-6-2015

Accepted: 23-6-2015