Pattern Discovery for Text Mining Measured by Levenshtein Edit Distance

Layla A. AL.hak
Department of Computer Science
College of Science, University of Diyala
Diyala , Iraq
laylaabcd3@gmail.com

Naji M. Sahib
Department of Computer Science
College of Science, University of Diyala
Diyala, Iraq
al.sehib@sciences.uodiyala.edu.iq

***Abstract*— The text mining techniques are utilized to extract interesting knowledge or information from the text documents. The process of finding accurate knowledge in text documents represents the main challenge for the users. Since many existing techniques of text mining encompassed term-based approaches, these techniques face the issues of synonymy and polysemy. Over the years, the researchers have discovered the assumption that pattern-based methodologies must provide better performance than the term based ones. This paper proposes an effective and innovative technique for pattern discovery that contains the pattern taxonomy model and Levenshtein edit distance algorithms for improving the efficiency of utilizing and updating the obtained patterns to find interesting and relevant information. This system shows that the accuracy of PTM is 98.09% for a specific dataset.**

Keyword* — *Text mining, Pattern Discovery, Levenshtein, PTM

I. INTRODUCTION

The amount of textual based information is rapidly accumulating which stored electronically on our computers or Web. Any computer (laptop or desktop) is capable of accommodating enormous data amounts because of the improvements in the storage devices. The process of information accumulating is very easy, but the process of obtaining relevant information when demanded may be difficult. Because the size of collections is continued to rise, the process of data structures construction which is used for facilitating the relevant information retrieval is becoming the main issue. Another important issue is the capability of extracting particular features or patterns for meeting specific needs of information [1].

The term of data mining is also called knowledge mining that is the significant extracting of inherent, formerly unknown and possibly beneficial information from data in the database. It has many techniques such as Decision tree classifier, Neural network, Genetic algorithm, Rule extraction [2].

Different applications like business management and market analysis can benefit from the utilization of information extraction from a huge amount of data [3]. Texts are the most popular means of formally information exchanging. Although it is difficult to extract beneficial information from these means, there is a requirement in this modernistic life to have a tool of business intelligence that is capable of extracting beneficial information as quickly as possible with a low cost [4].

The process of text mining or text data mining works on getting information with high quality from the processed text. This information is commonly acquired from the patterns and trends devising via means like learning of statistical pattern. Typically, the text mining includes; Firstly, structuring the input text; Secondly, extract patterns, and lastly, evaluate and interpret the output [5].

The text mining techniques are very useful to users for finding the desired knowledge from a massive data amount. So, it is extremely significant to retrieve efficient and relevant information for the users. Term-based approaches were utilized previously to provide these requirements. But these approaches have several drawbacks like polysemy and synonymy. Polysemy refers to a word which has several meanings, while synonymy refers to the words which have the same meaning. For overcoming these drawbacks, the phrase-based approaches were presented. But these developed approaches also have several drawbacks such as lower statistical properties to terms, the occurrence frequency of the phrases is minimal compared with the keywords, a considerable number of redundancy and noisy phrases. To defeat the issues of phrase-based approaches, pattern mining based approaches have been presented. Pattern discovery is used as an effective technique for knowledge discovery in many applications [6]. This paper focuses on the improvement of an efficient mining algorithm to efficiently extract and use the derived patterns and apply them to the domain of text mining. The pattern-based method is used in the proposed system.

Lots of researches in the field of text mining have concentrated on improving effective mining algorithms to discover various patterns from larger text documents. Hence, finding interesting and useful patterns remains an open issue. Within this field, the techniques of data mining can be utilized for finding a variety of text patterns, like frequent item sets, and sequential patterns. N. Zhong et al. (2012) [7], proposed a technique for discovering patterns to remove the issues of low frequency and misinterpretation of text mining.

978-1-7281-4377-4/19/$31.00 ©2019IEEE

The experiments were conducted on the topics of TREC and RCV1 data collection. This technique is also utilized to deploy pattern and evolve processes for improving the obtained patterns in text documents. B. Laxman and D. Sujatha (2013) [8] presented a technique to discover patterns and then compute the patterns specificities to evaluate the concept of weights as per their distribution in the obtained patterns. This technique works on updating patterns which show ambiguity that is a characteristic called pattern evolution. Patterns deploying and evolving are also used. The obtained results on the prototype application expose that the obtained result is very beneficial in the text data mining field. V. Aswini and S. K. Lavanya (2014) [9] presented a technique that utilized the pattern taxonomy model for discovering the patterns from a huge data amount and seeking for important patterns. This technique also includes the pattern evolving and deploying processes for improving the efficiency of utilizing and updating derived patterns to find important and relevant information. In the same context, S. D. Gupta and B. P. Vasgi (2015) [10] proposed a system that concentrates on performing a specific manner for deriving the pattern.

Vaishali Pansare (2016) [6] proposed an approach that uses the association rule mining based on the AprioriAll algorithm for discovering frequent patterns in text documents effectively. This proposed approach finds a solution to the misinterpretation and low-frequency issues. The obtained results have demonstrated that the time of execution needed by the algorithm is minimal than the time needed by the other compared algorithms. H. M. Mahedi Hasan et al. (2018) [11], presented a technique for key terms extraction that is depending on semantic relation. This approach is working on extracting a specified number of keywords from documents for identifying the main text content. The data set is collected from various sources like newspapers, books, journals, etcetera. To extract these keywords, several machine learning and statistical techniques have been. Support vector machine shows a better result than the other utilized techniques according to precision.

The aim of this paper is to discover interesting patterns from text document while these patterns contain accurate information about the document, and reduce the time required for finding the patterns which describe the content of the document, and increase the accuracy of information extracted from the text by using data mining techniques.

II. The Proposed System

The proposed system uses the pattern-based approach in which the discovered patterns are specified more than the entire text documents. The obtained patterns are a set of terms derived from the documents. The datasets that utilized in this proposed system is "Opinosis Opinion Dataset 1.0 - Documentation"(daraset1), and Reuter_50_50 Data Set (dataset2) which stored in a file of ".txt". The utilized dataset1 includes sentences extracted from reviews on a specific subject. The total size of dataset1 is 699 KB with different text lengths. There are fifty-one of subjects with one hundred sentences for each topic. The main idea of Opinosis Opinion Dataset is to generate a short summary from a large amount of text. Reuter_50_50 (dataset2) is a part of RCV1. 50 writers were selected. 50 writers with at least one subtopic of the CCAT. This dataset consists of 2500 text (50 texts per writer). The overall size of dataset2 is 7.29 MB. Fig. 1 shows the overall architecture of the proposed system.

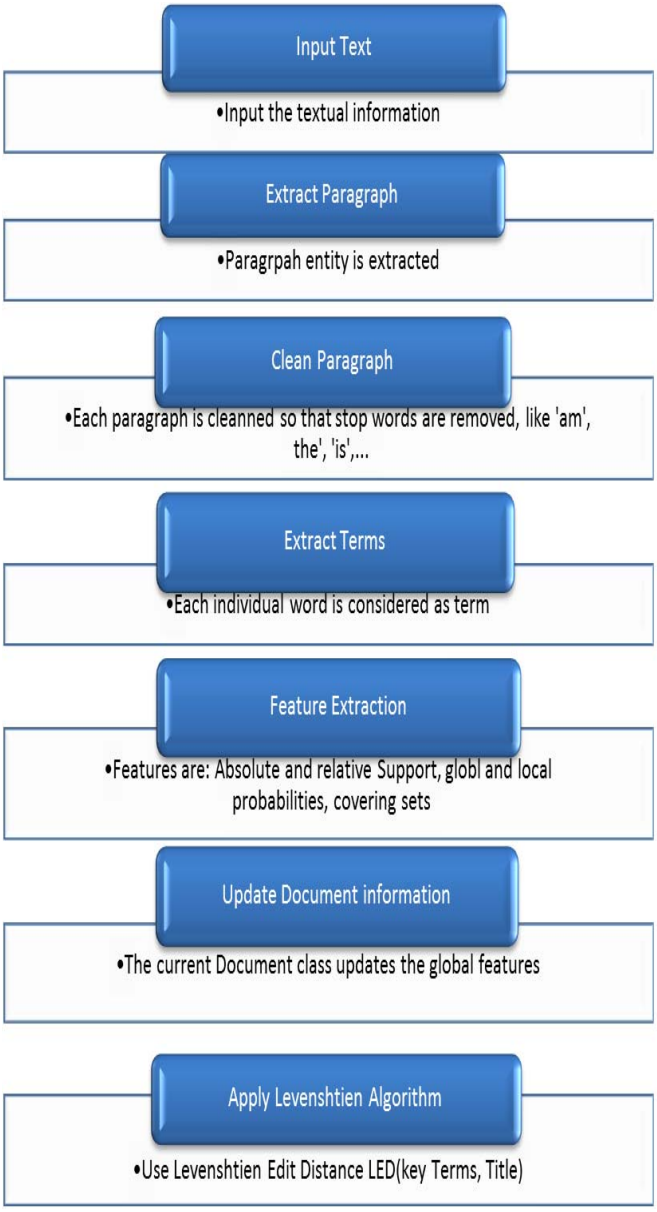

Fig. 1. System Architecture.

The proposed technique consists of several steps;

A. *Input Text step*

The input is entered via the end user. The input must be a text document.

B. *Extract paragraph Step*

In order to efficiently utilize the discovered pattern, this step is related to pattern taxonomy model (PTM). This model works on re-evaluating the patterns measures via deploying them into a common hypothesis space depending on their correlations in the taxonomies of the pattern. This results in high specificity patterns to the subject which can become

adequate and reasonable important values leads to an important development in the system efficiency.

PTM method firstly works on scanning the uploaded document and converting the entire document into a set of paragraphs which used as separate documents. After that, a process for extracting an amount of the set of terms from the obtained documents and terms form a specific pattern is done.

Algorithm 1: PTM Algorithm.

Input: a set of documents, a threshold
Output: a set of paragraphs
Begin
1: For each term T in document D
2: Assign threshold T_H
3: If SUPa ≥ T_H
4: Key Terms ←T
5: Else
6: Ignore T
7: End If
8: End For
9: Measure Accuracy

Here PTM algorithm compares the value of absolute support(sup_a) with the value of the threshold, and also the value of relative support (sup_r) and global probability and compute the accuracy of the feature with the high value.

A set of paragraphs are shown in Table I, to a given document d, here PS(d) = {dp1, dp2, …, dp6} all the redundant words are removed. Considering that min sup = 50%. Table II demonstrates '10' frequent patterns and their covering sets.

TABLE I. SET OF PARAGRAPHS.

Paragraph	Terms
dp_1	t_3 t_4
dp_2	t_1 t_2 t_3
dp_3	t_1 t_2 t_3 t_6
dp_4	t_1 t_2 t_3 t_6
dp_5	t_3 t_2 t_9 t_6
dp_6	t_5 t_4 t_3 t_2

TABLE II. FREQUENT PATTERNS AND COVERING SETS.

Frequent Pattern	Covering Set
$\{t_1,t_2,t_3\}$	$\{dp_1,dp_2,dp_3\}$
$\{t_2,t_4\}$	$\{dp_1,dp_2,dp_3\}$
$\{t_1,t_2\}$	$\{dp_1,dp_2,dp_3\}$
$\{t_2,t_3\}$	$\{dp_1,dp_2,dp_3\}$
$\{t_2\}$	$\{dp_1,dp_2,dp_3\}$
$\{t_1\}$	$\{dp_1,dp_2,dp_3\}$
$\{t_3,t_4\}$	$\{dp_1,dp_4,dp_2\}$
$\{t_3\}$	$\{dp_5,dp_3,dp_6\}$
$\{t_5\}$	$\{dp_2,dp_5,dp_6\}$
$\{t_4\}$	$\{dp_1,dp_2,dp_3,dp_4,dp_5\}$

There are some Terms used in Pattern Taxonomy Model:

1) Term Frequency: Term frequency (TF) is one of the main techniques for keyword extracting in which the word existence in the document is counting, for example, when TF is equal to "8", this means the text appeared "8" times in a document. Generally, if the TF is high, then the word is a significant one. TF can be calculated using the following equation:

$$TF = N_K / N \quad (1)$$

Where N_k is the ratio of the number of times a keyword k appears in a given document, to the total number of terms in that document N.

2) Threshold: The threshold is used for reducing the number of discovered patterns in a big document. These discovered patterns of minimum relative support will maximize the training burden.

C. Clean Paragraph Step

The purpose of this step is finding the key terms from the text document and improving the relationship between the word and document. This step includes:

1) Removing the Stop Word: Articles, prepositions, and pronouns are the most popular used words in the text documents that provide no meaning and can be considered as stop words. These words are not needful in the applications of text mining, therefore, these words will be eliminated. An instance of these stop words are 'a', 'an', 'the', 'in', 'and ', 'through', ' but ', ' themselves ', ' near '.... etc. This process requires 25 seconds to be completed.
2) Applying Porter Stemming: The process of stemming is used to minimize inflected or sometimes derived words to their root form or stem base. In order to get the root of the word, the porter stemming can be used. This process requires 25 seconds to be completed.

D. Extract Terms

After applying stop –word removal and porter stemmer**,** each individual word in each paragraph are considered as a term**.**

E. Feature Extraction

Feature extraction is an important pre-processing step which are used to represent texts**.** This process requires 1 minute and 34 seconds to be completed. The extracted features are:

1) Absolute Support (sup_a): sup_a is the number of occurrence of terms (X) in PS(d), i.e.

$$sup_a(X) = |X| \quad (2)$$

2) Relative Support (Sup_r): Sup_r is the fraction of paragraphs which includes the pattern, i.e.

$$Sup_r(X) = |X| / |PS(d)| \quad (3)$$

PS(d) are the number of paragraphs in the document. The term set X is known a frequent pattern when it's Sup_r(or) sup_a ≥ min-sup.

3) Global probability: It is the probability of the term existence in the document.

4) Local probability: It is the probability of the term existence in the paragraph.
5) Covering Set: It is the total number of paragraphs in which the term appeared.

F. Update Document information

In order to utilize the semantic information in the pattern taxonomy for improving the discovered patterns performance in text mining, the discovered patterns must be interpreted by summarizing them for accurately evaluating the term support.

This process will be done by using the deploy pattern algorithm. So, a term with a higher value of TF would be no meaning when it has not cited via some significant parts of documents.

Algorithm 2: Deploy Pattern Algorithm

```
Input:  A set of terms
Output: The most frequent terms
Begin
Step 1: For Terms in Document
Step 2: Sort Terms using Timsort Algorithm in ascending order.
Step 3: Sort Terms using Timsort Algorithm in descending order.
Step 4: For each Term T in Document
Step 5:  If GIobProb(T)> 0.0095
Step 6:  Add T to Pattern List
Step 7:  End if
Step 8:  End for
Step 9:  End for
```

After pattern deploying step the resulting terms are sorting according to the covering set by using Timsort algorithm. Timsort is a hybrid stable sorting algorithm, derived from merge and insertion sorts, constructed for doing a well performing on many real-world data types. This algorithm obtains the data subsequences which are formerly ordered and utilizes that knowledge for sorting the residue more effectively. This is accomplished via merge an identified subsequence, named a run, with existing runs till certain criteria are done. This process took 1 minute and 10 seconds to complete.

G. The Step of Applying Levenshtein Edit Distance algorithm (LED)

In this step, Levenshtein Edit Distance (LED) Algorithm is used. LED is a programming algorithm works on finding the matching between similar and the most frequent words. The Levenshtein distance is the minimum number of activities (insertion, deletion, substitution) required to transform one term into the other. LED takes the resulting terms with the title of the document and measures the similarity among them according to the following equation:

$$P = 1 - (K / K_{max}) \quad (4)$$

Where $K_{max} = \max(N, M)$, $K \geq 0$, $M > 0$, $N > 0$, $P \in (0,1)$

N, M – lengths of two terms respectively.

This process requires 99 seconds to calculate the accuracy of a document. In this system, the LED has been used to test the efficiency of PTM algorithm and obtain more accurate results for pattern discovery.

III. Experimental Results

In this proposed system, the utilized algorithms are implementing in java JDK 8u201 with Eclipse Java 2018-12 programming language by using a laptop computer. The experiments were performed on an Intel (R) Core (TM) i5-4210U CPU @1.70 GHz 2.40 GHz, 64-bit Operating System, and 8GB RAM. Every document undergoes different approaches for extracting the patterns. The approach efficiency is measured by utilizing the average accuracy:

Average accuracy =Total Accuracy/ Counter *100% (5)

Where Total accuracy is the total accuracy for each file Counter is the number of files.

Table III describes the accuracy of global probability, absolute support, and the relative support for different threshold values.

TABLE III. The average accuracy for different threshold values for Opinosis Opinion Dataset (dataset1).

Threshold	The average accuracy of global probability	The average accuracy of absolute support	The average accuracy of relative support
0.5	98.09%	73.06%	73.23%
1	94.02%	86.48%	86.14%
2	78.01%	93.28%	93.28%
3	55.19%	95.45%	95.62%
4	39.5%	96.43%	96.63%

Table III, shows that the average accuracy of global probability (in dataset1) when the threshold value =0.5 is higher than any average accuracy for different threshold values. Fig. 2 shows a comparison of threshold values Vs. features values for dataset1.

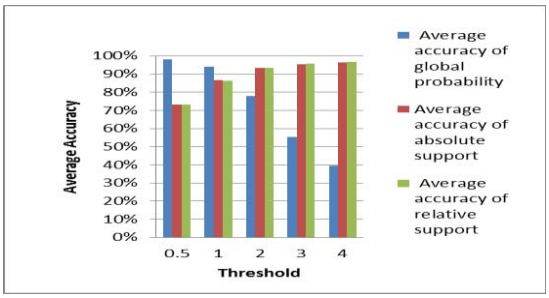

Fig. 2. Comparison of threshold values Vs. features values for dataset1

Table IV describes the average accuracy of global probability, absolute support, and the relative support for different threshold values.

TABLE IV. The average accuracy for different threshold values for Reuter_50_50 Data Set (dataset2)

Threshold	The average accuracy of global probability	The average accuracy of absolute support	The average accuracy of relative support
0.5	95.91%	11.28%	8.97%
1	86.17%	17.67%	18.13%
2	49.17%	41.34%	44.95%
3	30.30%	58.48%	61.31%
4	21.66%	68.57%	70.54%

Table IV shows that the average accuracy of global probability (in dataset2) when the threshold value =0.5 is higher than any average accuracy for different threshold values.

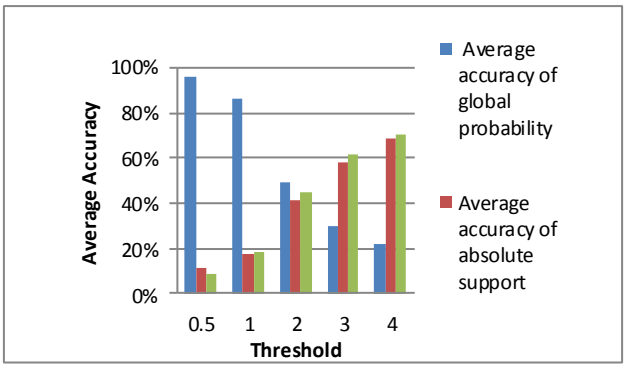

Fig. 3. Comparison of threshold values Vs. features values for dataset2.

So, in this proposed system, the perfect value of threshold is 0.5. Fig. 4 shows the absolute support when the threshold=0.5. Fig. 5 and Fig. 6 shows the relative support and global probability.

When comparing the proposed system with the other existing systems like the system presented by V. Aswini and S. K. Lavanya [9] which obtained an accuracy of 62% of precision and 82% of recall, and the system presented by H. M. M. Hasan et al. [11] which obtained an accuracy of 77.6% precision and 84.3% recall to measure the accuracy of the system, we found that the proposed system is more effective, where the obtained average accuracy is 98.09%.

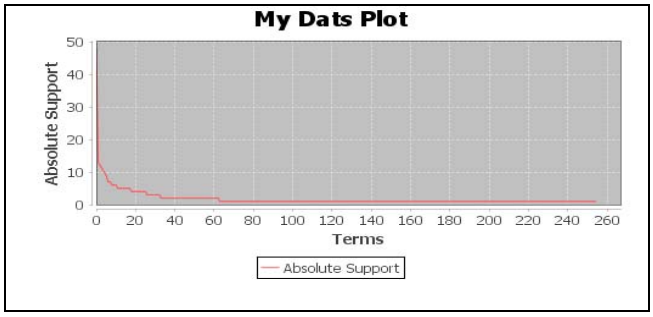

Fig.4. The absolute support for threshold=0.5.

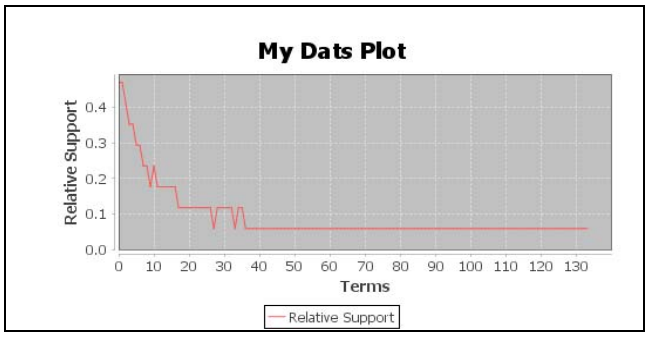

Fig.5. The relative support for threshold=0.5.

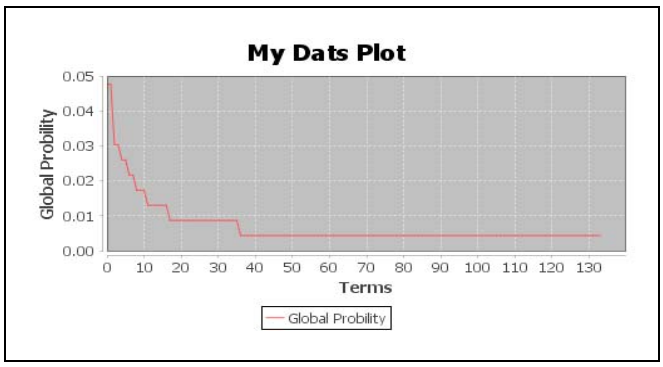

Fig.6. The global probability for threshold=0.5.

IV. CONCLUSIONS

Recently, several techniques of text mining have been presented to mining beneficial patterns related to what user's need. These techniques include association rule, sequential, maximum and closed pattern mining. Hence, utilizing the discovered pattern in the text mining domain is ineffective and not easy, because of beneficial long patterns with high specificity lack in support. This proposed system is focused on patterns discovering from a large amount of dataset.

In the proposed system, the model of pattern taxonomy, pattern deploy, and Levenshtein algorithm are utilized for pattern extracting from the documents. This system has been successfully performed and it is getting perfect results in text mining. Patterns are extracted with sufficient accuracy, depending on the documents.

In the future works, bi-gram terms will be extracted from the document, to investigate the accuracy as compared to the current uni-gram approach.

REFERENCES

[1] A. Akilan, "Text Mining: Challenges and Future Directions,"2015 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, 2015, pp.1679-1684.

[2] S. M. Ali and R.R Tuteja, "Data Mining Techniques," International Journal of Computer Science and Mobile Computing IJCSMC, Vol.3, No.4, April 2014, pp.879-883.

[3] M. Suganthy, K. Rupika, and J. S. Fransuva, "Text Mining for Pattern Identification," International Journal of Futuristic Science Engineering and Technology, Vol. 1, No. 3, March 2013.

[4] M. Inzalkar and J. Sharma, "A Survey on Text Mining-Techniques and Application," International Journal of Research in Science & Engineering, Special Issue: Techno-Xtreme 16, 2016.

[5] B. Shankaran, M. Patil, S. Suryawanshi, S. Mandhane, and S. S. Raskar, "A Novel Approach for Text Extraction using Effective Pattern Matching Technique," International Journal of Research in Engineering and Technology(IJRET), Vol. 04, No. 01, Jan 2015, pp. 269-272.

[6] V. Pansare, "Effective Pattern Identification Approach for Text Mining," International Journal of Computer Science and Information Technologies, Vol.7, 2016, pp.1826-1830.

[7] N. Zhong, Y. Li and S. Wu, "Effective Pattern Discovery for Text Mining," in IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No.1, Jan. 2012, pp. 30-44.

[8] B. Laxman and D. Sujatha, "Improved Method for Pattern Discovery in Text Mining," International Journal of Research in Engineering and Technology (IJRET), Vol.02, No. 10, Oct 2013, pp.574-578.

[9] V. Aswini and S. K. Lavanya, "Pattern Discovery for Text Mining," 2014 International Conference on Computation of Power,

Energy, Information and Communication (ICCPEIC), 2014, pp. 412-416.

[10] S. D. Gupta and B. P. Vasgi, “Implementation of Pattern Discovery to Retrieve Relevant Document using Text Mining,” 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, pp. 327-332, 2015.

[11] H. M. M. Hasan, F. Sanyal, D. Chaki and M. H. Ali, "An empirical study of important keyword extraction techniques from documents," 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), Aurangabad, 2017, pp. 91-94.