International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 10, ISSN No. 2455-2143, Pages 163-168
Published Online August - September 2016 in IJEAST (http://www.ijeast.com)

KEYWORD EXTRACTION FROM DESKTOP USING TEXT MINING TECHNIQUES

Dr. S.Vijayarani
Assistant Professor
Department of CSE,
Bharathiar University
Coimbatore

R.Janani
Ph.D.Research Scholar,
Department of CSE,
Bharathiar University
Coimbatore

S.Saranya
P.G Student
Department of CSE
Bharathiar University
Coimbatore

Abstract - **Information retrieval (IR) is used to identify the relevant documents in a large document database collection which is matching a user's query. The main goal of information retrieval system is to find the relevant information or a document that satisfies user information needs. The most important application of information retrieval system is search engine, example, Google search, Desktop search and Enterprise search, which identify the documents that are relevant to user queries. This research work focused on Desktop search. Desktop search is used to find information on the user's PC, which includes browser history, e-mail archives, text documents, sound files, images and video. The main objective of this research work is to retrieve the file name based on user given keyword from a collection of documents with various file extensions. In order to perform this task, this research work proposes a new keyword searching algorithm named as E-TFIDF. From the experimental results it is observed that the new keyword searching algorithm performance is better than existing TF-IDF algorithm.**

Keywords: Text Mining, TF-IDF, Desktop Search, keyword Search

I. INTRODUCTION

Information retrieval (IR) is determines the documents of an unstructured nature that satisfies an information need from a document collection. This system generally searches in collections of unstructured or semi-structured documents. The main applications of information retrieval systems are digital libraries, media search, search engine like desktop search, mobile search, and web search etc., [1]. This research work mainly focused on the desktop search to retrieve the file name based on user given keyword. Keyword extraction is tasked with the automatic identification of a collection of terms that best describe the subject of a document [2]. The main objective of this research work is to retrieve the file name based on user given keyword from a document collection with different file formats like .txt, .docx and .pdf.

This paper organized as follows, section II explains the literature survey and section III presents the methodology of this research work. Result and discussion given in Section IV and section V describes the conclusion of this research work.

II. RELATED WORK

Dipti S.Charjan, et.al. [4] Focused on developing efficient mining algorithm for discovering patterns from large data collection and search for useful and motivating patterns. In the field of text mining, pattern matching techniques can be used to discover various text patterns, such as frequent item sets, closed frequent item sets, co-occurring terms.

Bikash Mukhopadhyay et.al [5]. In this scenario, the volume of information is increased enormously, while the methods of retrieving that information remained relatively ineffective. The main source of difficulties in text retrieval research was natural language understanding barrier, which proved to be more challenging than anyone had predicted before. Fortunately it turned out that a lot of useful full-text analysis could be performed without a need to understand analyzed text contents, in a way similar to emerging data mining techniques.

Rafeeq Al-Hashemi [8]. The study introduces a sentence segmentation process method to make the extraction unit smaller than the original sentence extraction. The evaluation results show that the system achieves closer to the human constructed summaries (upper bound) at 20% summary rate. On the other hand, the system needs to improve readability of its summary output.

Beian Lott [9] methods have been used over the years, and new solutions are constantly being proposed to solve this complex problem. A broad overview of the common techniques and algorithms has not yet been explored. TF-IDF is one of the best-known and most commonly used keyword extraction algorithms currently in use when a document corpus is available. Several newer methods adapt TF-IDF for use as part of their process, and many others rely on the same fundamental concept as TF-IDF.

Zhang et al. [10] discusses the use of support vector machines for keyword extraction from documents using both the local and global context. There are number of techniques developed to use local and global context in keyword extraction. The other class of techniques used to enhance information retrieval uses concepts of semantic analysis such as ontology based similarity measures.

Anjali Ganesh Jivani [11] discussed that the purpose of stemming is to reduce different grammatical forms or word forms of a word like its noun, adjective, verb, adverb etc. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. This paper discusses different methods of stemming and their comparisons in terms of usage, advantages as well as limitations. The basic difference between stemming and lemmatization is also discussed.

III. Methodology

The main objective of this research work is to retrieve the file name based on user given keyword from a collection of documents. Based on the threshold value it will retrieve the file names from a document collection. In order to perform this task, this research work proposes a new algorithm named as Enhanced-TFIDF. The performance measures are used accuracy and time taken for searching the particular keyword. Figure 1 shows the system architecture of this research work.

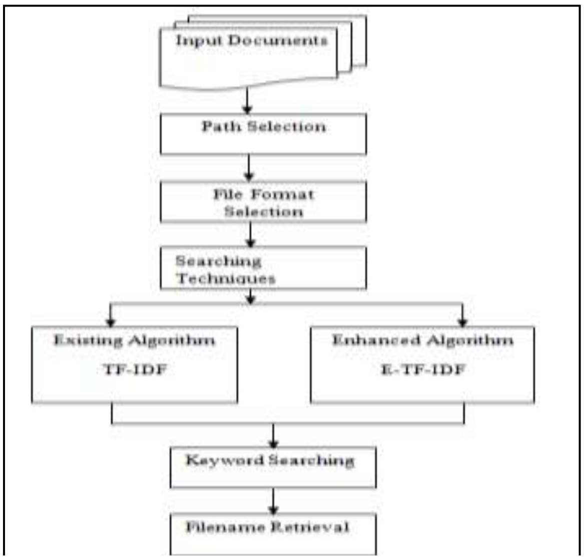


Figure 1: System Architecture

Existing Algorithm

A.TF-IDF (Term Frequency- Inverse Document Frequency)

TF-IDF is most commonly used keyword extraction algorithm in information retrieval. It is a numerical statistic that proposed to be a sign of how significant word is to the particular document in a collection of documents [12] [13]. All keyword extraction algorithms which make the use of a document collectiondepend on the weighted function. The tf-idf value increases proportionally to the number of times a word seems within the document; however is offset by the frequency of the word within the corpus, that helps to regulate for fact that some words appear a lot of often normally. It is frequently used as a weighting factor in information retrieval and text mining. Tf –idf can be used in various fields of text mining that includes text classification and text summarization.

TERM FREQUENCY (TF)

Term frequency (TF) is used to measure how frequently the particular term occurs in a document collection. Because every document has different size and it is possible that a particular term would appear more times in long documents than shorter documents [14]. The term frequency (t, d) is the simplest choice is to use the raw frequency of a term in a document. The number of times that term t occurs in document d.

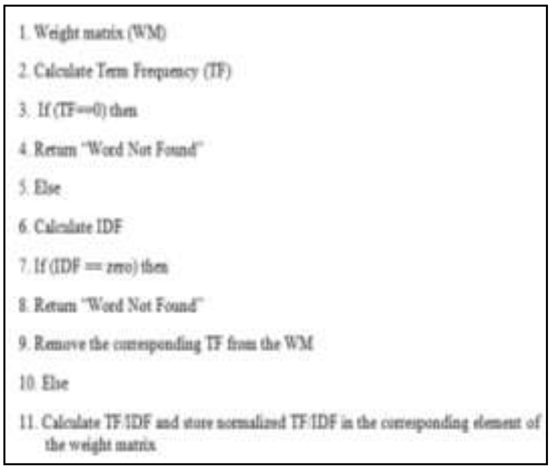
$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

INVERSE DOCUMENT FREQUENCY (IDF)

Inverse Document Frequency that measures however the term is important. Whereas computing the term frequency (TF), all terms should be considered as equally important terms. But it is acknowledged those certain terms, like "is", "the", “of” "that" etc.,, may appear a lot of times however have very little importance [17]. Thus need to weigh down the frequent terms while scale up the rare ones, by computing the subsequent IDF. It is the logarithmic scaled fraction of the documents that contain the word, obtained by dividing the total number of documents N by the number of documents d containing the term t, and then taking the logarithm of that quotient [18].

$$idf\ (t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Algorithm 1: Term Frequency- Inverse Document Frequency



```
1. Weight matrix (WM)
2. Calculate Term Frequency (TF)
3. If (TF==0) then
4. Return "Word Not Found"
5. Else
6. Calculate IDF
7. If (IDF == zero) then
8. Return "Word Not Found"
9. Remove the corresponding TF from the WM
10. Else
11. Calculate TF/IDF and store normalized TF/IDF in the corresponding element of
    the weight matrix
```

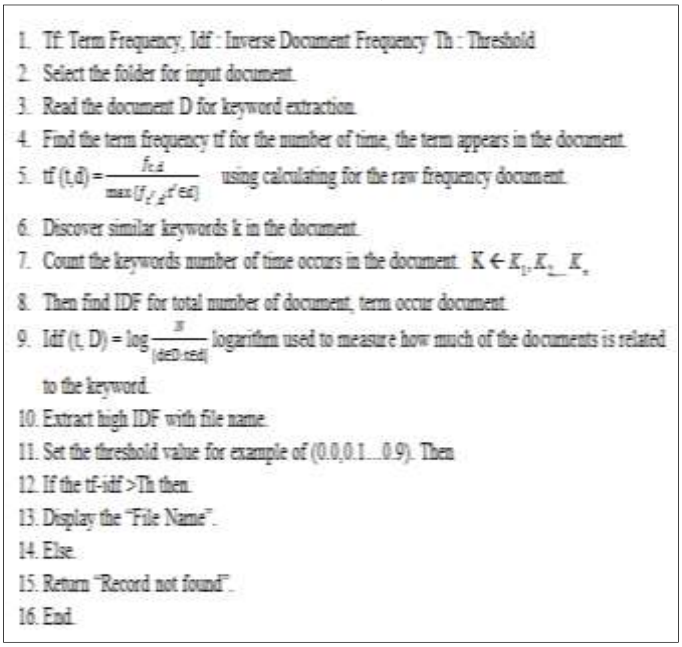
PROPOSED ALGORITHM

A. Enhanced Term Frequency Inverse Document Frequency (E-TFIDF)

The enhanced algorithm works as follows; first directory field is used to select the folder name which folder is user wants' to search for the keyword, by using the browse command. It may also include the subfolders for searching in the corresponding folder by select the check box. In selected folder it may contains a plenty of file format, Pattern field is used to filter those file format by the input. When click the ok button it may display the filtered file name in grid view or list box.

The keyword field is used to search inside the filtered file name for presences of the given keyword. If it returns yes then it may display the file name which files are contain the particular keyword. When click the Selection button it may produce the total length of the file and also it displays the total number of occurrence of the keyword in the given threshold range. Finally it will display the file name based on the given keyword and threshold value.

Algorithm 2: Enhanced Term Frequency- Inverse Document Frequency



```
1. Tf: Term Frequency, Idf : Inverse Document Frequency Th : Threshold
2. Select the folder for input document.
3. Read the document D for keyword extraction.
4. Find the term frequency tf for the number of time, the term appears in the document.
5. $tf(t,d) = \frac{f_{t,d}}{\max\{f_{t',d}: t' \in d\}}$   using calculating for the raw frequency document.
6. Discover similar keywords k in the document.
7. Count the keywords number of time occurs in the document. $K \leftarrow K_1, K_2 \ldots K_n$
8. Then find IDF for total number of document, term occur document.
9. $Idf(t, D) = \log \frac{N}{|\{d \in D: t \in d\}|}$ logarithm used to measure how much of the documents is related
   to the keyword.
10. Extract high IDF with file name.
11. Set the threshold value for example of (0.0,0.1...0.9). Then
12. If the tf-idf >Th then.
13. Display the "File Name".
14. Else.
15. Return "Record not found".
16. End.
```

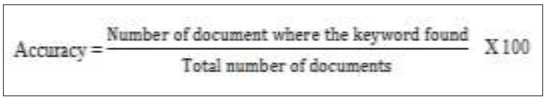
IV. Result and Discussion

In order to perform this analysis, the performance factors are search timeand relevancy for various types of file formats like .docx, .txt and .pdf. And the files on the desktop is taken as a dataset. For this analysis, the existing and enhanced keyword extracting algorithms were implemented by using Vb.net.

Example: The search word is "Mining", this algorithm used to search the particular word in all the documents and based on the threshold value it will retrieve the file name.

Search Time: It refers the time taken for searching the keyword within the document collection.

Relevancy: It refers the accuracy of the algorithm; the accuracy is calculated by using the formula as follows,



$$\text{Accuracy} = \frac{\text{Number of document where the keyword found}}{\text{Total number of documents}} \text{ X } 100$$

Table1and Table 2 shows that the details of input file like file name, the total number of words in a file and the size of a file.

Table 1: Input File Details

File Name	File Size (KB)	Total number of words in a file
Example1.txt	16	1879
Example2.docx	26	3200
Example3.pdf	580	5663

Table 2: Sample Input

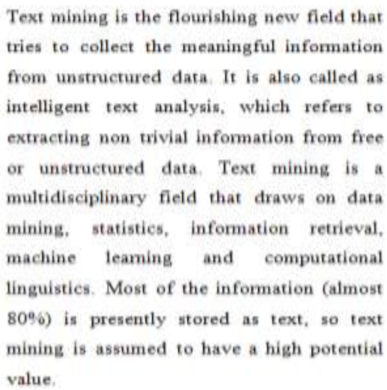
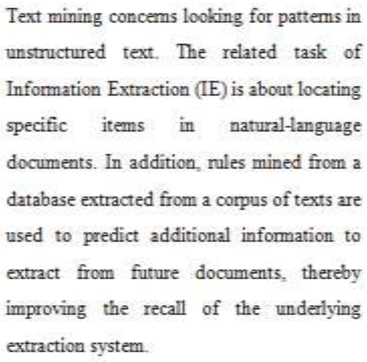
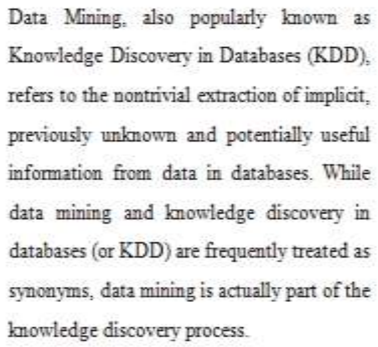
File Name	Sample Input
Example1 .txt	 Text mining is the flourishing new field that tries to collect the meaningful information from unstructured data. It is also called as intelligent text analysis, which refers to extracting non trivial information from free or unstructured data. Text mining is a multidisciplinary field that draws on data mining, statistics, information retrieval, machine learning and computational linguistics. Most of the information (almost 80%) is presently stored as text, so text mining is assumed to have a high potential value.
Example2 .docx	 Text mining concerns looking for patterns in unstructured text. The related task of Information Extraction (IE) is about locating specific items in natural-language documents. In addition, rules mined from a database extracted from a corpus of texts are used to predict additional information to extract from future documents, thereby improving the recall of the underlying extraction system.
Example3 .pdf	 Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

Table 3 shows the performance analysis of existing and enhanced TF-IDF algorithm for text files (Example1.txt). From this analysis the enhanced algorithm gives better accuracy when compared to existing algorithm.

Table 3: Performance analysis of TF-IDF and Enhanced TF-IDF Algorithm for text files

Algorithm	Time(ms)	Relevancy (%)
TF-IDF	31	65.97
E- TF-IDF	23	89.11

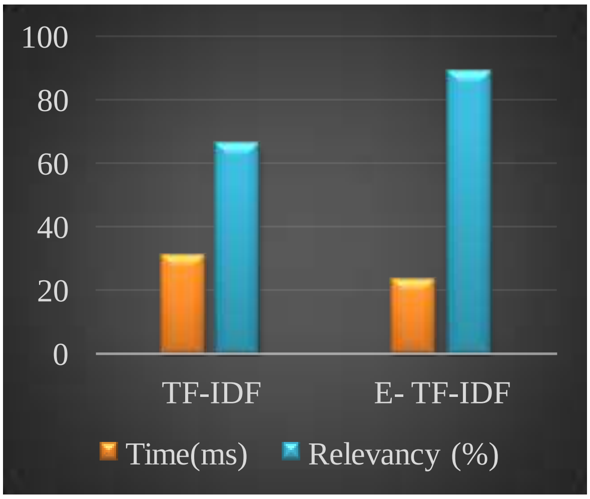


Figure 2: Performance analysis for text file

Table 4 shows the performance analysis of existing and enhanced TF-IDF algorithm for docx files (Example2.docx). From this analysis the enhanced algorithm gives better accuracy when compared to existing algorithm.

Table 4: Performance analysis of TF-IDF and Enhanced TF-IDF Algorithm for docx files

Algorithm	Time(ms)	Relevancy (%)
TF-IDF	42	84.74
E- TF-IDF	35	95.35

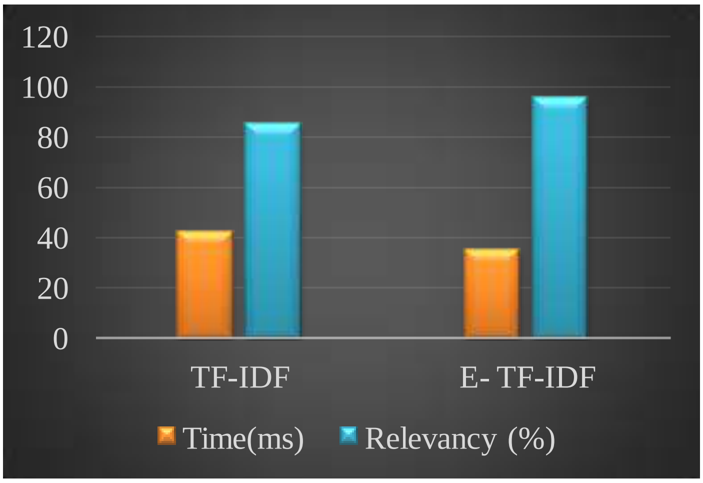


Figure 3: Performance analysis for docx file

Table 5 shows the performance analysis of existing and enhanced TF-IDF algorithm for pdf files (Example3.pdf). From this analysis the enhanced algorithm gives better accuracy when compared to existing algorithm.

Table 5: Performance analysis of TF-IDF and Enhanced TF-IDF Algorithm for pdf files

Algorithm	Time(ms)	Relevancy (%)
TF-IDF	61	50.89
E- TF-IDF	52	80.69

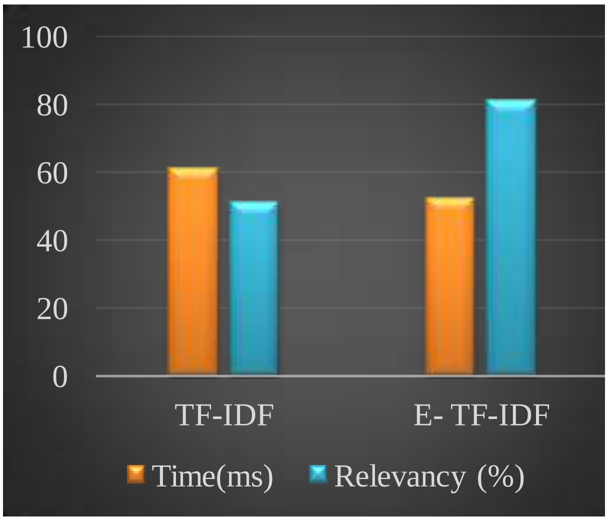


Figure 4: Performance analysis for pdf file

V. Conclusion

Information retrieval (IR) is used to recognize the important documents in a document collection which is matching a user's query. The main goal of information retrieval System is to find appropriate information that gratifies user information needs. The main objective this research work to exact the file name based on the keyword which is given by users. In order to perform this task this research work proposes a new algorithm. In the existing system, the text has been extracted but it produced lower accuracy, precision and recall performance. In the proposed system, the term frequency and inverse document frequency are calculated based on the threshold. The proposed system gives the higher performance and accuracy compared with existing system.

VI. References

[1] Un Yong Nahm, "Text Mining with Information Extraction", AAAI-2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, March 2002.

[2] Vashishta et.al,"Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 4, 2011.

[3] Xing Jiang and Ah-Hwee Tan,"Mining Ontological Knowledge from Domain-Specific Text Documents" Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05) 1550-4786/05 2005 IEEE.

[4] Miss Dipti S.Charjan and Prof. Mukesh A.Pund,"Pattern Discovery for Text Mining. Using Pattern Taxonomy", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 10- October 2013

[5] Bikash Mukhopadhyay et.al,"Data Mining Techniques for Information Retrieval", 2nd International CALIBER-2004, New Delhi, 11-13 February, 2004

[6] Ludovic Lebart et.al,"classification problems in text analysis and information retrieval".

[7] Vishakha D. Bhope and Sachin N. Deshmukh "Information Retrieval using Pattern Deploying and Pattern Evolving Method for Text Mining" International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3625-3629.

[8] Rafeeq A1- Hashemi "segment process method to extraction the original sentence".

[9] Beian Lott "TF-IDF is one of the best-known and most commonly used keyword extraction algorithms".

[10] Zhang et al support vector machines for keyword extraction from documents using both the local and global context.

[11] Anjali Ganesh Jivani , A Comparative Study of Stemming Algorithms, International Journal of Computer, Technology and Application, Volume 2, ISSN:2229-6093.

[12] Anette H, Karlgren J, Jonsson A, et al. Automatic Keyword Extraction Using Domain Knowledge[C]//Proceedings of Second International Conference on Computational Linguistics and Intelligent Text Processing. Mexico City: Springer- Verlag, 2001:472.

[13] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In

Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 2003

[14] S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: A System for Keyword-Based Search over Relational Databases. In ICDE, 2002

[15] Lang Zhou, Liang Zhang, Chong Feng, et al, Term Extraction Method Based on Word Frequency Distribution Change Statistic, J. Computer Science, 36 (2009) 177-180.

[16] Joachims and Thorsten, A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, No. CMU-CS-96-118.Carnegie-Mellon Univ Pittsburgh Pa Dept of Computer Science, 1996.

[17] G.Salton and C. Buckley.”Term -weighting approaches in automatic text retrieval”. Information Processing &Management,24.5.1998.

[18] H.Wu and R. Luk and K. Wong and K.Kwok.”Interpreting TF-IDF term weights as making relevance decisions”. ACM Transaction