International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 2, Mar-APR-2018, ISSN (Online): 2395-566X

A Survey on Privacy of Document Retrieval Different Techniques and Features Analysis

M.Tech. Scholar Nishu Kumari
Bansal institute of
science and technology
bhopal, India
nknishukumari@gmail.com

Asst.Prof. Ritu Ranjani Singh
Bansal institute of
science and technology
bhopal, India

***Abstract* – As the digital data increases on servers different researcher have focused on this field. From last few decades document are obtained from the various set gathered data from users, researcher, authors, etc. In this paper privacy of text document retrieval was detailed by various techniques of researchers. Here diverse attributes for the document privacy was explained with their necessities as attribute vary as per text study. Document classification algorithm are detailed with, text mining pre-processing working steps will remove noisy data from the initial dataset.**

***Keywords* – Supervised Classification, Text Mining, Text Feature, Text pre-processing**

I. INTRODUCTION

As the size of unstructured data in our world continues to increase, text mining tools that allow to sift through this information with ease will become more and more valuable. Text mining tools are beginning to be readily applied in the biomedical field, where the volume of information on a particular topic makes it impossible for a researcher to cover all the material, much less explore related texts[2].

Text mining methods can also be used by the government's intelligence and security agencies to try to piece together terrorist warnings and other security threats before they occur. Another area that is already benefiting from text mining tools is education. Students and educators can find more information relating to their topics at faster speeds than they can use traditional adhoc searching.

The new developments in text mining technology that go beyond simple searching methods are the key to information discovery and have a promising outlook for application all This one of a kind feature assumes an essential part in settling on issues that is related to money, social or different ramifications. Looking for second or third or numerous more opinions have fuelled the enthusiasm of analysts in the field of text mining.

Text Mining, also known as knowledge discovery (KD) from text, and document information mining (IM), refers to the procedure of extracting fascinating information from very large text quantity for the purposes of determining knowledge[4].

It is an interdisciplinary field involving IR, understanding text, extraction of information, clustering, classification, linkage of concept, visualization, database knowledge, machine learning (ML), and DM. Search engine is the most well known Information Retrieval tool.

Application of Text Mining techniques to Information Retrieval can improve the precision of retrieval systems by filtering relevant documents for the given search query[4].
Whole paper was divide into five section where first was introduction which include importance and requirement of document analysis.

While second include various features of text mining for document retreival. Third section provide brief related work done by different authors. Fourth section provide text pre-processing methods ie. Stopt word, stemming.

II. FEATURES OF TEXT MINING

1. Title Feature
The word in sentence that likewise happens in title gives high score. This is controlled by checking the quantity of matches between the substance word in a sentence and word in the title. In [4] ascertain the score for this element which is the proportion of number of words in the sentence that happen in the title over the quantity of words in the title.

2. Sentence Length
This feature is valuable to sift through short sentence, for example, datelines and writer names usually found in the news articles the short sentences are not anticipated that would has a place with the synopsis. In [5] utilize the length of sentence, which is the proportion of the quantity

of words happening in the sentence over the words happening in the longest sentence of the records.

3. Term Weight
The recurrence of the term event with records has been utilized for computing the significance of sentence. The score of a sentence can be ascertained as the aggregate of the score of words the sentences. The score of essential score wi of word I can be computed by customary tf.idf technique.

4.Sentence position
Regardless of whether it is the initial 5 sentence in the section, sentence position in content gives the significance of the sentences. This feature can include a few things, for example, the situation of the sentence in the records, area and the section, and so forth, proposed the principal sentence of most elevated positioning. The score for this feature in [6] consider the initial 5 sentence in the section.

5. Sentence to sentence similarity
This component is a similitude between sentences for each sentence S , the likeness amongst S and each other sentence is figured by the cosine closeness measure with a subsequent incentive in the vicinity of 0 and 1 [6]. The term Weight wi and wj of term t to n term in sentences Si and Sj are spoken to as the vector. The similitude of each sentence combine is ascertained in light of closeness.

III. CLASSIFICATION TECHNIQUES

The primary work of blog level opinion grouping is to recognize the polarities of UGC (user generated content). Two kind of grouping systems have been utilized as a part of record level document arrangement, regulated strategy and unsupervised technique.

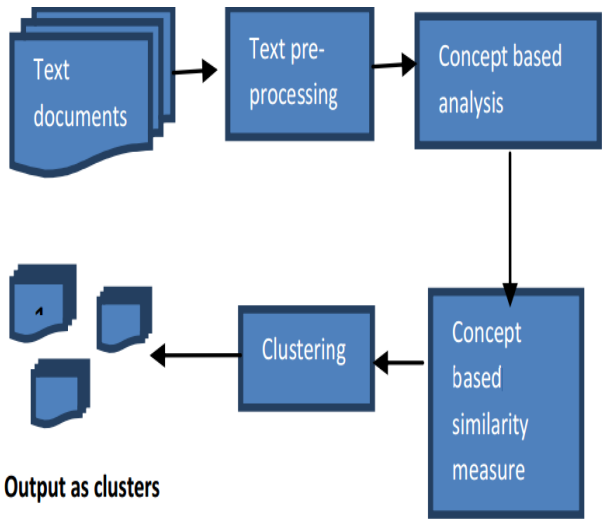

Fig. 1 Text document clustering approach.

1. Supervised Methods
Document arrangement can be detailed as a directed [4] learning issue with four classes, positive, negative, impartial and productive. Client produced content generally are utilized as training and testing information. Any current regulated learning methods can be utilized to document grouping, for example, support vector machines (SVM) and naïve Bayes.

2. Sentence level Document Classification
The assignment of ordering a sentence as subjective or target is regularly called subjectivity grouping. The subsequent subjective sentences are a likewise delegated communicating positive or negative opinion, which is called sentence-level assessment characterization.

In the sentence level text document examination, the extremity of each sentence is computed. This is like an archive level feeling examination yet done at a sentence level [3]. It accept each sentence contains an assessment for one substance and angle, and a portion of the sentences may not be obstinate (objective). The subjective sentences contain assessment words which help in deciding the opinion about the element.

A two phase surmising is improved the situation each sentence: to start with, each sentence is named subjective or target and afterward the extremity of every one of the subjective sentences are construed. There might be unpredictable sentences additionally in the stubborn content. In such cases, sentence level opinion characterization isn't valuable.

3. Aspect level Document Classification
In an ordinary obstinate blog, the writer composes both positive and negative parts of the substance, in spite of the fact that the general document on the element might be sure or negative. Archive and sentence text document grouping does not give such data.

To get these subtle elements, this work have to go to the perspective level. It accept that a record contains document on a few substances and their angles. Viewpoint level characterization requires disclosure of these substances, perspectives, and opinions for every one of them.

IV. RELATED WORK

In [2] Ziqi Wang, et al. discussed the well-known dictionary matching algorithm called as Aho-Corasick algorithm. The AC tree is a trie with “failure links”, on which the Aho Corasick string matching algorithm can be executed.

The Aho-Corasick algorithm is a well-known dictionary matching algorithm which can quickly locate the elements

of a finite set of strings within an input string. The time complexity of the algorithm is of linear order in the length of input string plus the number of matched entries [15].

In [3] Keyvanpour and Tavoli proposed feature weighted technique for improving performance of document image retrieval system. Feature weighting is an approach that approximates the optimal degree of influence of individual features of document images. This method weights the feature using coefficient of multiple correlations.

In [4] Pirlo et. al presented a method for layout based document image retrieval using dynamic time warping for commercially designed documents. Morphological operations are used for extracting grid based structural components from the document image. Random transform is used for description of the layouts and dynamic time warping is used for document indexing.

In [5] author propose schemes to deal with privacy preserving ranked multi-keyword search in a multi-owner model (PRMSM). To enable cloud servers to perform secure search without knowing the actual data of both keywords and trapdoors, this work systematically construct a novel secure search protocol.

To rank the search results and preserve the privacy of relevance scores between keywords and files, this work propose a novel additive order and privacy preserving function family. To prevent the attackers from eavesdropping secret keys and pretending to be legal data users submitting searches, this work propose a novel dynamic secret key generation protocol and a new data user authentication protocol.

In [6] propose a procedure where a query is not (primarily) based on single terms, but on a set of reference documents. Compared to the problem of determining concrete key terms for a query it is rather easy for analysts to manually compile a collection of 'paradigmatic' documents which reflect topics or manner of speech matching their research objective. Retrieval with such a reference collection is then performed in three steps:

- Extraction of a set of key terms from the reference collection, called dictionary. Terms in the dictionary is ranked by weight to reflect difference in their importance for describing an analysis objective.
- Extraction of co-occurrence data from the reference collection as well as from an additional generic corpus representative of a given language.
- Scoring relevancy of each document on the basis of dictionary and co-occurrence data to create a ranked list of documents.

V.TEXT PREPROCESSING

As blogs are accumulations of passages. Sections are gathering of sentences. While sentences are accumulation of words. So entire preprocessing center around word in the archive with no accentuations. So in pre-handling of archive there are two normal advances initially is stop word evacuation, and second is stem word expulsion [3].

1. Stop List Removals

As sentence is outline with number of words yet some of those words are simply use to construct an appropriate sentence despite the fact that it doesn't make any data in the sentence. So distinguishing proof of those words at that point evacuating is term as Stop word expulsion. So a rundown of words is store by the analyst which help in recognizing of stop words [5, 6].

This expulsion of stop words help in diminish the execution time of the calculation, in the meantime boisterous words which not give any productive data is likewise evacuated. Stop words resemble {a, the, for, an, of, and, etc.}. So text document is transform into collection of words which is then compare with these words and then each match word is removed from the document.

In order to understand this assume an sentence {India is a great country in the world} then after pre-processing it become {India, great, country, world} while stop words {is, a, in, the} in the sentence are removed.

2. Stem Word Removal

In this words which are almost similar in prefix are replace by one word. This can be said collection of words share same word is term as stem [10]. So there occurrence in the document make same effect but while processing in text mining algorithm it make different so update each word from the collection into single word is done in this stem word removal pre-processing step. Let us assume an collection of words for better understanding of this work. Collection of word is {play, plays, playing} then replaces each with word {play}.

VI. CONCLUSIONS

As the writing work of different document, articles from organization, press media, institutes are increasing day by day. Then publishing their work is also increase which is done by most of the news paper, organizations. Here paper has cover an important issue of text document retrieval. Various techniques with their required features are discussed in detailed. Here paper related work of researchers done in this field. So it can be concluded that one strong algorithm is required that can effectively classify and retrieve document on the basis of author document.

REFERENCES

[1]. Dandan Jiang1, Xiangfeng Luo1, Junyu Xuan, And Zheng Xu .“Document Computing for the News Event Based. on the Social Media Big Data”. Digital Object Identifier 10.1109/ACCESS.2016.2607218 IEEE Acess 2017

[2]. Ziqi Wang, Gu Xu, Hang Li, and Ming Zhang,” A Probabilistic Approach to String Transformation”,published in IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 5,pp. 1063-1075, May 2014.

[3]. M. Keyvanpour and R. Tavoli, “Feature Weighting for Improving Document Image Retrieval System Performance,” International Journal of Computer science, vol. 9, pp.125-130, 2012.

[4]. Giuseppe Pirlo, Michela Chimienti, Michele Dassisti, Donato Impedovo, Angelo Galiano, “Layout Based Document Retrieval System by Radon Transform Using Dynamic Time Warping,” Image Analysis and Processing –ICIAP 2013, Lecture Notes in Computer Science vol. 8156, pp 61-70, 2013.

[5]. Wei Zhang, Yaping Lin,, Sheng Xiao, JieWu, Fellow, IEEE, and Siwang Zhou. “Privacy Preserving Ranked Multi-Keyword Search for Multiple Data Owners in Cloud Computing”. Ieee Transactions On Computers, Vol. 65, No. 5, MAY 2016.

[6]. Gregor Wiedemann and Andreas Niekler. “Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries”. arXiv:1707.03217vol 11 Jul 2017

[7]. Fangzhao Wu, Yongfeng Huang, Yangqiu Song, Shixia Liu,” Towards building a high quality micro blog-specific Chinese document lexicon”, Decision Support Systems-2016.

[8]. V.K. Singh, R. Piryani, A. Uddin, P. Waila," Document Analysis of Movie Reviews", conference on IEEE-2013

[9]. Selma Ayşe Özel. Esra Saraç “ Web Page Classification Using Firefly Optimization “, 978-1-4799-0661-1/13/$31.00 ©2013 IEEE.

[10]. Yuefeng Li, Ning Zhong, and Sheng-Tang Wu “Effective Pattern Discovery for Text Mining”. IEEE transaction on knowledge and data engineering Vol. 24 no. 1 Jan 2012.

[11]. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, And Moch Arif Bijaksana “Relevance Feature Discovery For Text Mining” . Ieee Transactions On Knowledge And Data Engineering, Vol. 27, No. 6, June 2015