International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Evaluation of Literature Survey Classification System based on TF-IDF and Stemming Technique using RNN Algorithm

Miss.Kshitija G. Deshmukh and Prof.Dr.(Mrs) S. A. Itkar

Department of Computer Engineering, PES Modern college of Engineering, Pune

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Various research papers are published online and offline, and as new research fields have been continuously created, users have tons of trouble in finding and categorizing their attention-grabbing research papers. The proposed system extract the abstracts of each paper then, it removes the stop word by preprocessing method and it also removes the suffix of word by using stemming technique. The stemming technique is used to reduce the high dimensionality of the feature space. Then, the Recurrent Neural Network (RNN) algorithm is applied to classify the research papers with similar subjects, based on the Term frequency-inverse document frequency (TF-IDF) values of each paper.

Keywords: *Research Paper Classification, Recurrent Neural Network, Categorization, Stemming Technique.*

Introduction

Number of research papers has been published with the increasing advance of computer and information technologies, which makes it difficult for users to search their interesting research papers for a selected domain [1].Therefore particular number of research paper are classified with same domain so that users can find their interesting research papers easily.

The most of the time used evaluation of the type of a big range of research papers is administered on massive-scale computing machines with none concept on big data properties. The relation of the papers to be analyzed and categorized is extremely complicated, successfully classify research papers with the same domains in terms of contents. Hence needed a preprocessing method for large scale of research paper, it gets an accurate and quick classification.

To classify a large range of papers into papers with similar domain, In prious work solution was given using the term frequency-inverse document frequency (TF-IDF) [24],Stemming Technique [5]. The proposed system first extract abstract from research paper, then preprocessing is done.After the preprocessing method stemming technique remove the prefix, suffix and infix of the word. It uses the TF-IDF scheme to extract words from the abstract of papers. Then RNN rule is used for classifying research papers with similar domains, supported the Term Frequency – inverse document frequency (TF-IDF) values of every paper.

The remaining paper is standardized follows. Literature survey is done in section 2. Proposed system in Section 3. Section 4 explains the algorithm of the proposed system. Section 5 Discusses about performance parameter and experimental result. Section 6 conclusion of proposed system.

Literature Survey

This section shortly reviews the literature on various paper classification technique implemented till date. Document classification has direct, familiar with the paper classification of this paper. It is a problem that assigns a document to at least one or a lot of predefined categories according to specific contents. The descriptive application areas of document classification are as followed: The K-means algorithm is applied to classify the total papers into research papers with similar subjects, using the Term frequency-inverse document frequency (TF-IDF) values of every paper [1][5].

These neural networks known as recurrent as a result of this step are moved out for each input.As these neural networks take into account the previous word throughout predicting, it acts like a memory storage unit that stores it for a brief amount of time [7]. A recurrent structure of capture discourse data as well as feasible once learning word representations, which may introduce significantly less noise compared to ancient window-based neural net-works [8].

An artificial Datasets such as News 20, Reuters, email, and analysis papers on completely different topics. Term Frequency-Inverse Document Frequency formula is employed along with fuzzy K-means and hierarchical formula[2].TFIDF method and framework for text classification. The framework allows classification in

line with varied parameter, measure and analysis of results[3][4].
The main of porter stemmer uses suffix denudation in English language. This stemmer could be a linear step stemmer. Specifically, it's 5 steps applying rules inside every step [11]. Inside every step, if a suffix rule matched to a word, then the conditions connected to its rule are tested on what would be the ensuing stem, if that suffix was removed, within the means outlined by the rule[9][10].

Proposed Methodology

The research classification during this paper consists of 4 main processes.(Fig. 1): (1) crawling, (2) Stemming and Data Management (3) TF-IDF (4) Classification. This section describes a system flowchart for our paper classification system.

A. Architecture

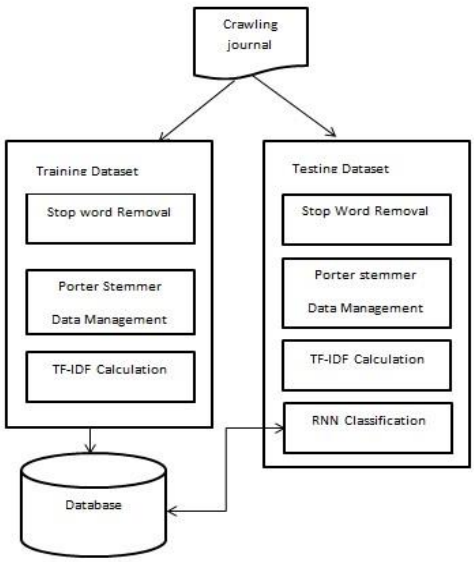

Fig. 1. System Architecture

Detailed flows for the proposed system flowchart shown in Fig.1 as follows:
Step1: It automatically collects keywords and abstracts data of the IEEE paper.
Step2: From abstract of paper remove stop word.
Step3: Then apply stemming technique on abstract for Remove suffix of word.
Step4: It calculates number of occurrences of words in the abstract of each paper.
Step5: It calculates a TF value for each keyword occur in abstract.
Step6: : It calculates an IDF value for each keywords occur in abstract
Step7: It calculates a TF-IDF value for each keyword using the values obtained by 5 and 6
Step8: It classify the papers into papers with a similar do mains, based on the RNN algorithm.

B. Data Preprocessing

Data preprocessing suggests that changing unstructured information into structured information. Given a matter supply containing differing kinds of document (different formats, language formatting) the primary action that ought to text processing.

1. Removal of stop word and symbol

Preprocessor processes words by removing Symbols removal, Stop words removal. All the symbols are removed in preprocessing step and a stop list is a list of commonly repeated features which appears in every abstract. The common features such as it, he, she and conjunctions such as and, or, but etc. are to be removed because they do not have the result of the categorization process [13].
Stop words are typically one set of words. It means that completely different for various varieties of applications. As an example, a stop word list will contain.

- Determiners: the, a, an, another
- Coordinating conjunctions: for, an, nor, but, or, yet, so
- Prepositions: in, under, towards, before

C. Stemming Technique

In research paper classification porter affix removal stemmer is to remove the ending of the word keeping 1st n letters, i.e. to truncate a word up to ordinal character and take away the remainder. Many affix removal stemming algorithms was developed by the researchers[13].
Original porter stemmer rule consists of solely five steps. Every step applied the principles, and condition additionally concerned. If the rule is properly accepted, the suffixes are mechanically removed in step with the condition, and also the next step performed. The principles and conditions finish at the resultant stem [10].
Step 1: Remove suffix from word "SSES" i.e. (crosses Cross) Step 2: Remove suffix from word "IES" i.e. (studiesstudi) Step 3: Remove suffix from word "SS" i.e. (caress- care) Step 4: Remove suffix from word "S" i.e. (dogs- dog) Step 5: Remove suffix from word "EED" i.e. (agreed- agree)
In our system each step includes again suffix for improved efficiency of stemming. It includes suffix for removal i.e.
'ing','ity','ly','tion','al','lize','ness','ment' and so on.

1. Managing Data

Constructs the keyword lexicon mistreatment the abstraction knowledge and keyword knowledge crawled in crawl step and saves it to the HDFS. In order to method various Keywords merely and expeditiously, this paper categorizes many keywords with similar meanings into one representative keyword.In the existing system construct 1394 representative keyword from total keywords of all abstract and build a keyword lexicon of those representative keyword[1].

1.1 Keyword Data

A keyword is a word that serves as a key, as to the meaning of another word, a sentence, passage, or the like. Key-word data are data having word and each word represent as meaningful keyword.
In our system get word from the abstract. Each word represents a keyword. Example "Word is recorded" in this sentence 'word' and 'record' is the keyword.

1.2 Keyword Dictionary
Collecting all keyword from data, create dictionary contain all keyword of data. In proposing system construct keywords from the total keywords of all abstracts and make a Keyword dictionary of these representative keywords.
1.3 Abstract Data and Research Data
A paper title, successive most a part of the papers that uses square measure seem to read is that the abstract. That is, the user tends to browse a paper abstract so as to catch the analysis direction and outline before reading all con-tents within the paper [1].Abstract data in a research paper, usually in one paragraph of 300 words, the major aspects of the entire paper in a prescribed sequence that include the overall purpose of the study and the research problem, the basic design of the study.
D. TF-IDF Model
TF-IDF technique is employed that eliminates the most common terms and extracts solely most relevant terms from the corpus[2].Term frequency-inverse document frequency (TF-IDF) could be a numerical data points technique that permit the determination of weight every for every term (or word) in each document[3].
1. Term Frequency (TF)
The method computes the quantity of repetitions of a word (term) within the document [3]. In our system term frequency is calculated occurrence each word in one abstract. Term frequency calculates as follows:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where,
ni,j:The number of occurrences of word ti in document dj and k nk,j:The total number of occurrences of words in document dj
K and D:The number of keywords and documents (i.e., papers), respectively.
2 Inverse Document Frequency
The inverse document frequency issue is incorporated that diminishes the load of terms that occur terribly often within the document set and will increase the load of terms that occur seldom. Inverse document frequency is a measure of how much information the word delivers, i.e. if it's common or rare across all documents.

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

where,
N : Total number of document in corpus N=|*D*| d^{P}*D* : *t*^{P}*d* : *numberofdocumentswherethetermtappears.*
3 Document Frequency
While the TF means that the amount of occurrences of every keyword during a document, the DF means that what number times every keyword seems within the assortment of documents. The DF is calculated by dividing the whole range of documents by the amount of documents that contain a selected keyword. It's outlined as [1]

$$\text{Dfij} = \frac{|D|}{|\{d \in D : t \in d\}|}$$

Where,
|*D*| : The total number of documents d ^{P}*D* : *t*^{P}*d* : *Thenumberofdocumentsthatkeywordtoccurs.*
4 TF-IDF TF-IDF (Term Frequency –Inverse Document Frequency) is employed to convert documents into structured format [7]. TF-IDF methodology determines the ratio of words in specific document through associates in the inverse proportion of words over the complete document corpus [8].This calculation determines however relevant a given word is in an exceedingly particular document. Word that square measure common in an exceedingly single or a tiny, low cluster of document tend to own higher TF-IDF number than common words [9].
*TF-IDF = TF * IDF*
IV. ALGORITHM
A. Recurrent Neural Network Algorithm
Recurrent Neural Networks are one amongst the foremost common neural network employed in tongue process because of its promising results. The applications of RNN in language models accommodate 2 main approaches [7].A repeated neural network (RNN) [Elman, 1990] is in a position to process a sequence of absolute length by recursively applying a transition operate to its internal hidden state vector ht of the input sequence. The activation of the hidden state of ht at time step t is computed as operating f of the present input symbol Xt and therefore the previous hidden state ht-1[13].

$$\text{ht} = \sum_{f(ht-1,Xt)}^{0} \quad \begin{matrix} t=0 \\ otherwise \end{matrix}$$

In proposing system RNN algorithm used for classifying the research paper their interesting domain.

The similarity vector will return the current weight of test object with all training instances. Classification has done with respective weight factor.Classification has done with respective weight factor. It will assign the label according to maximum weight generated by the algorithm. Final phase works for base classification. It provides a sub class categorization. Finally, similarity score will classify each bucket into the respective domain.

B. Comparative Study

This paper compared the proposed algorithm to existing algorithm. K-mean Classification - Higher time and space complexity stores all the instance, Noisy features degrades the classification accuracy, RNN - Ability to better capture the contextual information[6][7].

Sr. No.	Parameter	K means	RNN
1	Type of Algorithm	Unsupervised	Deep learning
2	Basic	*K-means* clustering is to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.	A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph
3	Classification	Classification on the basis of k cluster	Classification on basis of hidden neural network layer
4	Learning method	It good learning method Without supervision	It better learning method as compare to K mean because It working on deep learner method

Fig. 2. Proposed Algorithm Compare with Existing Algorithm

C. Hardware and Software Requirements

Hardware Requirements:

1. Processor - Pentium IV 2.4 GHz.
2. Hard Disk - 40 GB
3. Floppy Drive - 44 Mb
4. Monitor - 15 VGA Color

Software Requirements:

1. Services Web Based
2. IDE - Eclipse Oxygen
3. Front End - .jsp
4. Back End Servlet/Data Base(MYSQL)

Result And Discussion

In this section describe the result and discussion of proposed system. In result, we show the each step of the system. First the data divide into training data set and testing dataset.Given result shows the training module.

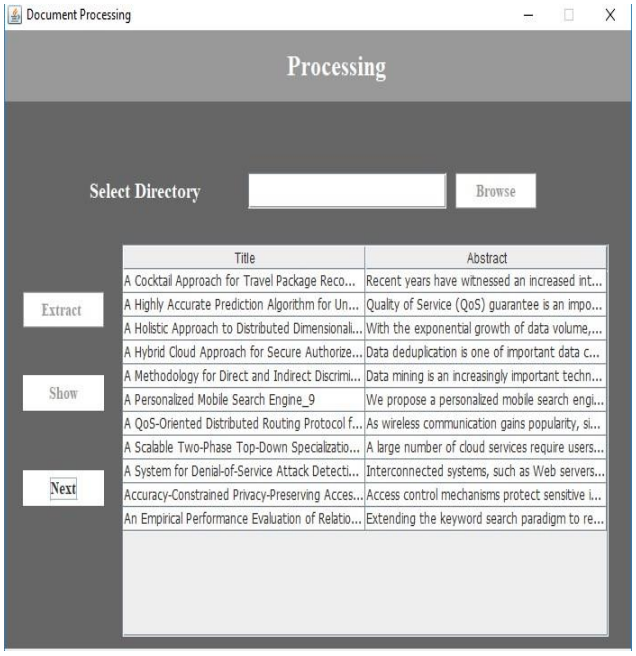

Fig. 3. Crawling Abstract from Papers

In above fig. 3, firstly select the train directory containing IEEE journal after the click on extract it extract abstract for each journal paper in the training dataset. In below fig 4 in this after crawling the abstract remove all stop word.

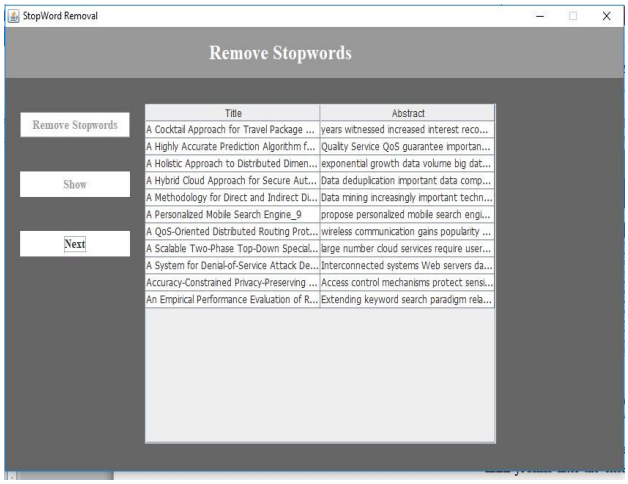

Fig. 4. Removal of Stop Word

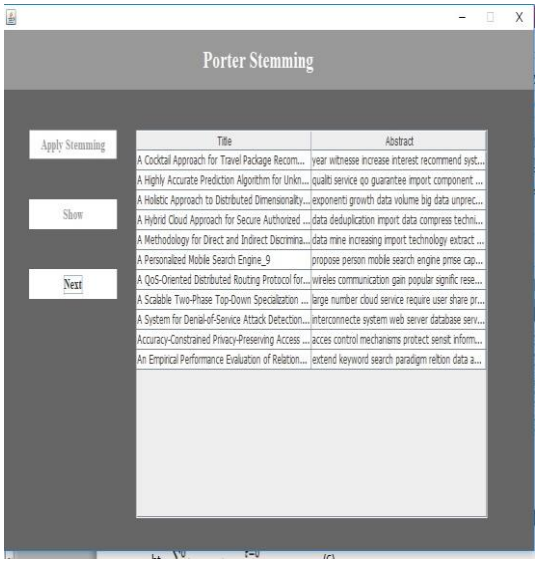

Fig. 5. Porter Stemmer

In figure 5 porter stemmer it remove the suffix of the word. After clicking on ' apply stemming', stemming algorithm remove suffix using its rule.

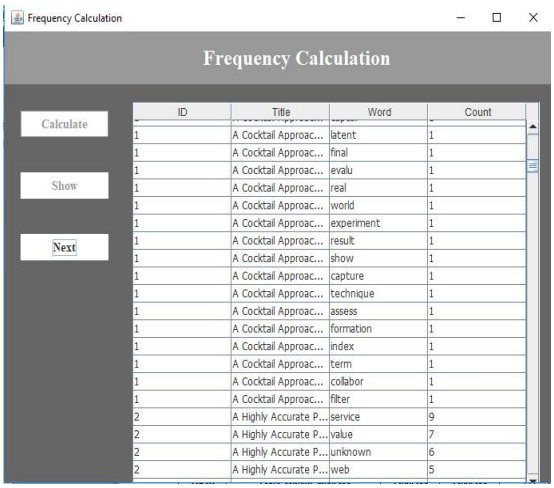

Fig. 6. TF Calculation

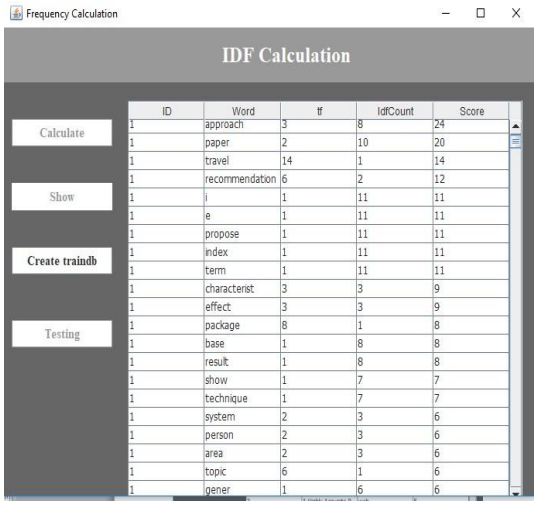

Fig. 7. Calculation Inverse Document Frequency and (Term Frequency * Inverse Document Frequency) Calculation

In above fig. 6 and 7 calculation of TF and IDF for each term. Number of occurrences of each word in fig 6. And in fig 7 total number of occurrences of each word in whole abstract, calculate the TF-IDF.In fig 8 all the train dataset stored in the database. Each abstract classify in their interesting domain it stored in a database.

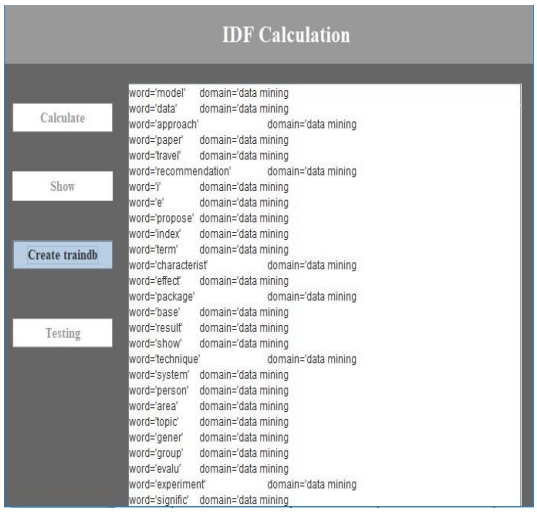

Fig. 8. Create Train Database

Conclusion

The intention of this paper is to research paper classification based on TF-IDF, stemming technique and learning algorithm for classification. Our studies in space conclude that reduce the high dimensionality of feature space using porter stemming technique. Classification systems can classify research papers in advance by both of keywords and stemming with the support of high-performance computing techniques.

The experimental results showed that the proposed system can classify the papers with similar subjects according to the keywords extracted from the abstracts of papers. Classified research papers will be applied to search the papers within users' interesting research areas, fast and efficiently. More efficient classifiers for research paper datasets.

References

[1]. Sang-Woon Kim and Joon-Min Gil," Research paper classification systems based on TF-IDF and LDA schemes" Kim and Gil Hum. Cent.
[2]. Comput. Inf. Sci. (2019).
[3]. Prafulla Bafna, Dhanya Pramod, Anagha Vaidya, "Document Clustering: TF-IDF approach", IEEE 2016.
[4]. Bruno Trstenjak,Sasa Mikac, Dzenana Donko, "KNN with TF-IDF Based Framework for Text Categorization", Procedia Engineering 69,science Direct (2014) 1356 – 1364
[5]. Juan Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries", JURAMOS@EDEN.RUTGERS.EDU
[6]. Vairaprakash Gurusamy, S.Kannan, K.Nandhini,"Performance Analysis: Stemming Algorithm for the English Language ",IJSRD - International
[7]. Journal for Scientific Research Development Vol. 5, Issue 05, 2017 ISSN (online): 2321-0613
[8]. Pema Gurung and Rupali Wagh ,"A study on Topic Identification using K means clustering algorithm: Big vs. Small Documents" ,Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 2 (2017) pp. 221-233.
[9]. D. Yogeshwaran1, Dr. N. Yuvaraj ," Text Classification using Recurrent Neural Network in Quora" International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 02 Feb 2019.
[10]. Pengfei Liu ,Xipeng Qiu, Xuanjing Huang,"Recurrent Neural Network for Text Classification with Multi-Task Learning", Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)
[11]. Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao,"Recurrent Convolutional Neural Networks for Text Classification", Proceedings of the TwentyNinth AAAI Conference on Artificial Intelligence
[12]. Hanumanthappa M and Narayana Swamy M,"Language Independent Categorization of Documents Based on the Domain", Advances in Natural and Applied Sciences, 9(6) Special 2015, Pages: 545-548
[13]. Mrs. R. Jayanthi , Ms. C. Jeevitha,"An Approach for Effective Text Pre-Processing Using Improved Porters Stemming Algorithm", IJISET - International Journal of Innovative Science, Engineering Technology, Vol. 2 Issue 7, July 2015.
[14]. Jashanjot Kaur, Preetpal Kaur Buttar,"A Systematic Review on Stopword Removal Algorithms", International Journal on Future Revolution in
[15]. Computer Science Communication Engineering April 2018 ISSN: 24544248 Volume: 4 Issue: 4
[16]. Ms. Anjali Ganesh Jivani, "A Comparative Study of Stemming Algorithms" Int. J. Comp. Tech. Appl. IJCTA NOV-DEC 2011, Vol 2 (6), 1930-1938.