Asian Journal of Information Technology 15 (18): 3430-3440, 2016
ISSN: 1682-3915

Frequency Based Modified Term Weighting Method for Text Classification

[1]M. Santhanakumar, [1]C. Christopher Columbus and [2]K. Jayapriya
[1]Department of Computer Science and Engineering, PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu, India
[2] Department of Computer Science and Information Technology
Nadar Saraswathi College of Arts and Science
Theni, Tamil Nadu, India

Abstract: Due to the huge amount of data on the World Wide Web (WWW), it is very important that the users can access the related details without losing any valuable information. Term weighting based on the user query plays a vital role in Information Retrieval (IR). Term Frequency-Inverse Document Frequency (TF-IDF) is one of the repeatedly used term weighting method which assigns weights based on the occurrences of a term in a document. This paper proposes a Modified Term Frequency (MTF) using multi term occurrences in a document. In the proposed work, the weight is assigned to the documents based on the occurrences of the co-terms in a document and it is classified to find the accuracy using three different classifiers such as Support Vector Machine (SVM), Decision Tree (DT) and K- Nearest Neighbor (KNN). The experimental result shows that the classification accuracy and other performance measures such as precision, recall and f-measure of the propose work outperforms the some of the existing other term weighting methods.

Key words: Term frequency, inverse document frequency, term weighting, classification, India

INTRODUCTION

In the current decade, the quantity and availability of information have been growing in a tremendous rate on the web. The Web has massive amount of semi-structured and unstructured data. Since, web users have some difficulties like retrieving irrelevant data while accessing information through internet. Web Mining (WM) tool can help the users to discover potential information from a huge amount of raw data. The tasks of WM are mainly classified into three categories such as Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM). The process of retrieving the relevant document based on a keyword is known as WCM which uses Information Retrieval (IR) and Natural Language Processing (NLP) as the two technologies for retrieving the web document (Santhanakumar and Columbus, 2015a). Since, web search becomes the root of IR, reasonable amount of effort has been made to access the relevant documents that the user needs. On the internet, most of the information is in the form of web text (Liangtu and Xiaoming, 2007).

Vector Space Model (VSM) is one of the good methods to express semi-structured and unstructured information of web. In VSM, Term Frequency-Inverse Document Frequency (TF-IDF) is a commonly used method for calculating term weighting based on the occurrences of a term within a document (Lee *et al*., 1997). Term Frequency (TF) and Inverse Document Frequency (IDF) are calculated by finding the number of occurrences of a term in a single document and finding the number of documents containing that term respectively (Salton, 1989). Even though TF-IDF is a commonly used term weighting method, it has some drawbacks such as it handles only single term and considers multiple occurrences of the terms as a single entry (Xia and Chai, 2011). To overcome such limitations, number of modified term weighting methods have been proposed in literature (Sabbah *et al*., 2016; Santhanakumar and Columbus, 2015b). In some other cases, text is represented statistically based on the lexical, syntactic, bag of words and n-grams features (Sabbah *et al*., 2016). However, the text classification by these methods causes low performance because of its inability to understand the semantic meaning of a user query (Choi *et al*., 2014). So the classical TF-IDF and the statistical term weighting methods are not adequate for text classification. This paper, proposes a modified term weighting method based on the classical TF-IDF. The weighting of the term has to be assigned depends on the co-occurrences of a term.

Correspondig Author: M. Santhanakumar, Department of Computer Science and Engineering, SN College of Engineering and Technology, Tirunelveli, Tamil Nadu, India

The experiment results are compared with other term weighting methods such as TF, TF-IDF and Entropy. Multi class Support Vector Machine (MSVM) classifier is used for text categorization.

Literature review: Before finding the weight of a term, some of the pre-processing techniques such as tokenization (Zulkifeli *et al*., 2012), stopword removal and stemming (Porter, 1980) are to be applied to remove unwanted data and find the root word of the terms inside the document. For text categorization Boolean weighting, word frequency weighting, TF-IDF and entropy weighting are some of the term weighting methods (Liu and Peng, 2014) in IR. Various term weighting methods have been implemented based on the single term weighting method such as TF, DF and TF-IDF (Sabbah *et al*., 2016).

A genetic algorithm based (Zaefarian *et al*., 2006) modified TF-IDF has been developed and proved to have a better recall value than the classical TF-IDF. To assign weight for multi-party spoken languages used by speakers in a meeting, an algorithm named SU-IDF (Murray and Renals, 2007) was developed. It is also extended from classical TF-IDF. In this method, the precision value is increased by 10 points by assigning higher weight to the rare terms in a document. The term weighting method Term Frequency-Relative Frequency (TF-RF) has been implemented (Lan *et al*., 2009) which assigns higher weight to frequently distributed term in the positive category. A modified term weighting method (Fang *et al*., 2011) has been implemented based on analytical constraints and diagnostic test. In this method, the document has more query terms and also distinct query terms have been assigned as a higher weighted one. Wang *et al*. (2010, 2015) have developed an improved term weighting scheme, which assigns weight based on the term distributed inside a single class and multiple classes. A modified TF-IDF weighting scheme (Xia and Chai, 2011) has been implemented based on the term uniformly distributed and widely appeared inside a document. Term Frequency-Inverse Sentence Frequency (TF-ISF) has been developed (Doko *et al*., 2013), which assigns weight by considering the background of a sentence like previous and next of the current sentence. The queries such as short queries and long queries are combined together (Paik, 2013) to develop a novel TF-IDF scheme which produces better recall value than the classical TF-IDF. In multiterm, a new measure called Term Frequency-Information Content (TF-IC) has been used to prioritize the crawled content and link (Pesaranghader *et al*., 2013).

An improved term weighting approach (Gautam and Kumar, 2013) has been developed by considering the important and rare term with higher weight even, if it has a low frequency in a document. A novel term weighting method has been implemented (Wang and Zhang, 2010) based on the inverse category frequency. A method called Document Frequency-Inverse Corpus Frequency (DE-ICF) (Goswami and Kamath, 2014) has been developed and it depends on the number of times the user visited a document and number of corpus containing that document. Lexical cohesion based topic segmentation method (Bouchekif *et al*., 2014) has been implemented without providing any external information about the data. A novel term weighting method (Carmel *et al*., 2014) has been developed based on synthetic information of the query term available in a document. A new weighting scheme called Term Frequency Inverse Positive Negative Document frequency (TFIPNDF) (Liu and Peng, 2014) has been developed to represent the importance of term in positive and negative category based on the distribution of the terms. A hybrid method for feature selection process has been developed (Liu *et al*., 2014) by combining term frequency and document frequency information such as Optimal Document Frequency based Feature Selection (ODFFS) and normal Term Frequency based discriminative power measure and comprehensively measure Feature Selection (TFFS). A modified TF-IDF method (Sabbah and Selamat, 2014) has been implemented to categorize the dark web content. Term weight has to be assigned based on the number of unique terms in the document collection. The accuracy of this scheme compared with Entropy (Selamat and Omatu, 2003), Glasgow (Sanderson and Ruthven, 1996) and TF-RF (Lan *et al*., 2006; Zaefarian et al., 2006). A new feature selection method Document frequency and Term frequency combined Feature Selection (DTFS) has been developed (Wang *et al*., 2015) for e-mail classification.

Multiword weighting scheme has been developed based on the properties of semantic and statistical quality. The experiments were conducted on Chinese and English corpus and compared with traditional TF-IDF and Latent Semantic Indexing (LSI). Three different approaches (Attia *et al*., 2010) have been used in heterogeneous data resources to extort the multiword terms. Three new word association term extraction techniques (Huo, 2012) namely statistical association measure, smoothed probabilities of N-grams and normalized sequence probabilities have been developed for term weighting. Multi Term Adjacency Keywords Order (MTAKO) is a vector space model (Raj, 2012) developed to improve the relevancy of search results based on two assumptions. B-ranking scheme (Ricardo, 2013) used to

extract the multiword terms based on their relevance from the different collection of documents without using dictionary.

Although many term weighting methods are implemented based on single term and multi term, selecting a better feature selection remains still a challenging task in web mining. The motivation behind this work, term weighting is an origin of either text classification of clustering. TF-IDF is the most frequent method used for term weighting. However, several modified term weighting methods were implemented to improve the performance of classification and clustering. This leads to propose a new term weighting method based on the occurrences of co-terms in a document in order to achieve higher classification than other methods.

MATERIALS AND METHODS

Corpora: There are 20 Newsgroup dataset accumulations used to confirm the recovery exactness and adequacy of proposed work. This dataset contains Usenet articles Ken Lang gathered from 20 diverse Newsgroups recorded in Table 1 and 19,997 archives in it.

For every category, it holds about 1,000 documents. In this paper, just 10 categories from this dataset are used to evaluate the performance of the recommended system. The documents from these datasets are further splitted into test set and training set. Test set will be used to produce those test queries randomly and training set will be used to test the trials. The number of documents on every category under the training dataset have been indicated in Table 2.

The documents in every group are pre-processed as tokenizing, stop word exclusion and stemming. Porter stemmer calculation may be used to displace the negligible terms from those documents. The various Query Vector (QV) are randomly formulated using the test dataset. The QV is denoted in Eq.1.

$$QV = qt_1, qt_2, \ldots, qt_n \quad (1)$$

where, n is the number about exceptional terms in the accumulation. Furthermore, t_i (i=1, 2...n) depicts each expression of an inquiry vector.

Preprocessing: Preprocessing is one of the key phases in text classification. To improve the efficiency of text classification, preprocessing is necessary to be performed on web documents, removing the irrelevant data like images, audio, numbers, symbols,

Table 1: Categories in 20 Newsgroup dataset

Categories	News grop
Alt.atheism	Rec.sport.hockey
Comp.graphics	Sci.crypt
Comp.os.ms-windows.misc	Sci.electronics
Comp.sys.ibm.pc.hardware	Sci.med
Comp.sys.mac.hardware	Sci.space
Comp.windows.x	Soc.religion.christian
Misc.forsale	Talk.politics.guns
Rec.autos	Talk.politics.mideast
Rec.motorcycles	Talk.politics.misc
Rec.sport.baseball	Talk.religion.misc

Table 2: The number of documents in each topic from 20 Newsgroup training set

Category	No. of Documents
Alt.atheism	480
Comp. graphics	584
Comp.os. ms-windows.misc	587
Misc.forsale	585
Rec. motorcycles	598
Rec. sport. hockey	600
Sci. electronics	591
Sci.space	593
Soc. religion. christian	599
Talk. politics. mideast	397

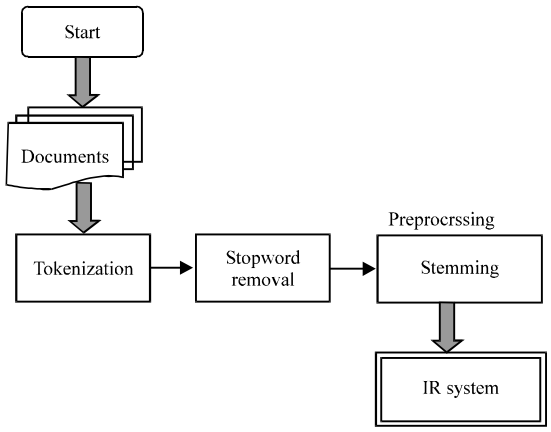

Fig. 1: Process of preprocessing

stopwords, etc and finding root word of the terms. Figure 1 describes the process of preprocessing.

Tokenization: It (Zulkifeli *et al*., 2012) is the critical process of segregating and removing separate whitespaces, punctuation symbols, alphabetic strings and alphanumeric symbols. The identified numeric characters, words and other characters are called as tokens.

Stopword removal: The output of the tokenization process is passed to stopword removal phase as shown in Fig. 1. Here the English terms like if, what, was and, or, as, of, etc., which are ineffective for IR are removed from the documents. The size of the indexed files are also reduced which improves the overall efficiency and effectiveness of the classification process.

Stemming: It is the process of finding the root word of the terms present in the documents. For example, the term "study" is obtained from the term "studying" by removing suffix of the term. This also improves the efficiency of IR model. For this purpose Porter stemming algorithm (Porter, 1980) is used in this research.

Term Weighting schemes: Term weighting schemes like TF, Document Frequency (DF), IDF are frequently used methods in IR, Text Classification (TC) and Web Classification (WC) (Sabbah *et al*., 2016). In addition to that Entropy (Selamat and Omatu, 2003) and Glasgow (Sanderson and Ruthven, 1996) are another two methods used in term weighting. However, Term Variance (TV), Term Strength (TS), Chi-Square (CHI), Information Gain (IG), Gini Index (GI), Mutual Information (MI) and Balanced Term Weighting Scheme (BTWS) are the some other term weighting methods proposed and implemented in IR, TC and WC (Sabbah *et al*., 2016). This section describes some of the existing term weighting methods which are used for comparison with the proposed work.

Term Frequency (TF). This is a simple method to determine the relevant document to the user query. It also neglects the documents that do not contain the terms in a user query (Sabbah *et al*., 2016). TF of the term is calculated by counting number of occurrences of the term in a document. Since, TF variation depends on the length of a document which it is normalized using the length of that document. Normally TF is denoted by $tf_{t,d}$, where, t and d represent query term and document, respectively. The normalized TF is described in Eq. 2:

$$tf_{t,d} = \frac{n_{t,d}}{\sum_k n_{k,d}} \quad (2)$$

where, n and k denote the number of occurrences of the term t and length of the dth document, respectively. Inverse Document Frequency (IDF). Inverse Document Frequency is a global term weighting method. It supports the term, which occurs rarely in the document set rather than the term frequently occurs (Sabbah *et al*., 2016). IDF is defined by the Eq. 3.

$$idf_t = \log \frac{|M|}{|\{d : t_t \in d\}|} \quad (3)$$

Here, M denotes the total number of documents in a corpus and the denominator describes the number of documents containing the term t_t. IDF has only the positive values because the denominator is always lesser than or equal to M (Sabbah *et al*., 2016). Term Frequency-Inverse Document Frequency (TF-IDF). In many information retrieval applications, classical TF-IDF is used for weighting the term (Chiang *et al*., 2008). Because, TF-IDF is an outstanding method that considers less frequent term as most significant one in a document (Sabbah *et al*., 2016). TF-IDF is the dot product of TF and IDF of the term. The TF-IDF of the term is denoted by $TF\text{-}IDF_{t,d}$ and it is described in Eq. 4:

$$TF - IDF_{t,d} = tf_{t,d} \times idf_t \quad (4)$$

Entropy: Entropy is another weighting method which works based on probabilistic analysis and information theory. In entropy, most frequent term in less number of documents considered as important term that distributes over the collection of documents (Sabbah *et al*., 2016). Entropy weighting is divided into local and global weighting which is described in Eq. 5 and 6:

$$"L"_{"t,d"} = "\begin{cases} "1+""log""tf"_{"t,d"}, ("tf")_{"t,d"} > 0 \\ "@"0, "("tf")_{"t,d"} = 0" \end{cases} \quad (5)$$

$$G_1 = \frac{1 + \sum_{d=1}^{n} \frac{tf_{t,d}}{Ft} \log\left(\frac{tf_{t,d}}{Ft} + 1\right)}{\log N} \quad (6)$$

Where, F_t denotes the frequency of the term t in the entire document. Entropy is the dot product of local and global term weighting that is described in Eq. 7:

$$w_{t,d} = L_{t,d} \times G_t \quad (7)$$

Modified Term Frequency (MTF): Term weighting is used for shaping the significant level of the term for a document based on the variables such as term frequency, length and term specificity. In most of the IR applications, TF-IDF is the frequently used term weighting method. Even though, the performance of TF-IDF is well in most of the situations, it has some drawbacks (Keh *et al*., 2010).

So variety of term weighting methods have been developed (Santhanakumar and Columbus, 2015a, b). This study proposes a new term weighting method based on the classical TF-IDF. It includes the proportion of the sum of frequencies of common keyterms in each document to the total frequency of common keyterms in all the documents identified from a corpus. If there is no common keyterms in the documents of a corpus, then it calculates

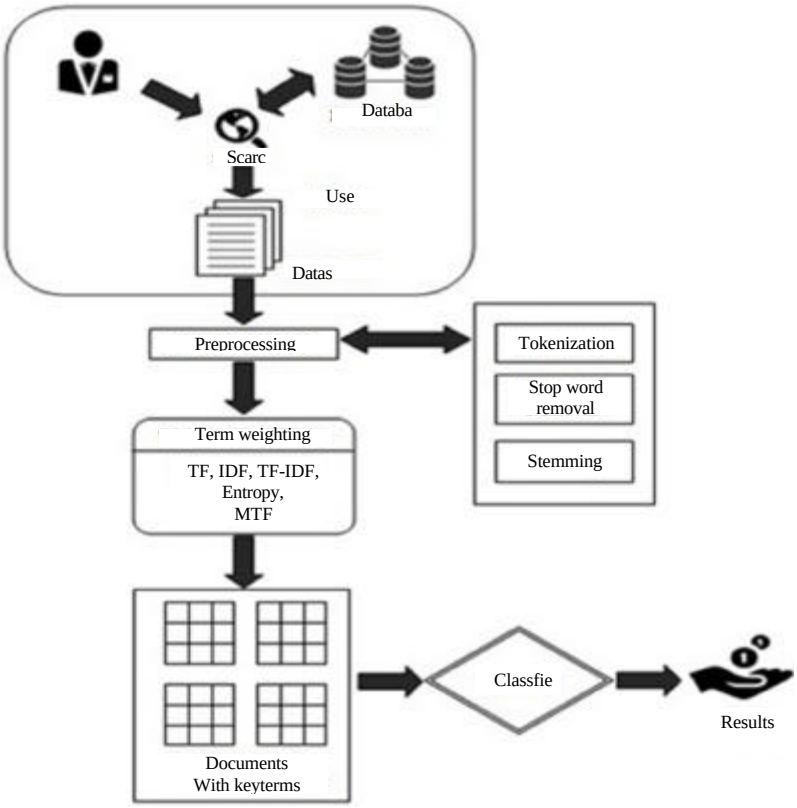

Fig. 2: Experimental architecture of MTF

the frequency of that single keyterm in each document and total frequency of that keyterm in all the documents of that corpus. The proposed work is denoted by MTF. Figure 2 illustrates the experimental architecture of the proposed work.

Let us consider, N number of keywords and the corpus C contains the D number of documents. In the proposed work, let SM be the set of documents containing i^{th} keyterm k_i and then weight of the document sd_j is calculated as defined in Eq. 8:

$$MTF_{t,d} = \frac{n_{k_i sm_j} + \sum n_{kc} sm_j}{\sum n_{ki} SM + \sum n_{kc} SM} \quad (8)$$

Where:

sm_j = The jth document in SM

k_c = The occurrence of common keyterms in all the documents of SM

Let us consider the document set $D = \{d_1, d_2, d_3, ..., d_n\}$ where, N is the total number of documents in a corpus and $SM = \{sm_1, sm_2, sm_3, ..., sm_n\}$ be the document set that contains the keyterms $K = \{k_1, k_2, k_3, ..., k_n\}$. The following steps describe the working principle of the proposed research:

- For each document d_j of corpus D search for the keyterm k_i in a corpus
- Then, find if all other keyterms occur in the same documents SM where k_i occurs
- If so, find the sum of frequencies for each keyterm in each document. Else, calculate the frequency of that individual keyterm k_i in all documents if $k_i \in d_j$
- Continue this process to compute total number of frequencies of all documents if more than one keyterms commonly occur in the same document d_j
- Find the final term weight by the product of individual term weight and classical TF-IDF. The highest weighted document is preferred first by the user
- Sort all the documents in descending order according to the weight and select the top weighted documents

Table 3 describe the way of assigning weights to the document based on the proposed work. For example from Table 3 by considering the keyterm k_1, the documents d1, d2, d4 and d5 containing that term k_1. Similarly the

Table 3: Classification performance measurement

Weighting Methods	Classifiers	Alt.atheism	Comp. Graphics	Comp.os.ms-windows.misc	Misc. forsale	Rec. Motorcycles	rec.sport. hockey	Sci.Electronics	Sci.space	Soc.religion. christian	Talk.politics. mideast
TF	DT	65.04	60.96	61.96	68.86	73.23	75.00	68.47	67.15	63.78	56.12
	KNN	76.62	73.45	75.78	80.98	84.52	87.47	77.45	79.43	76.13	69.98
	SVM	82.33	79.18	76.50	82.45	86.27	90.67	80.81	81.29	82.16	72.67
TF-IDF	DT	74.25	74.78	69.17	73.13	75.48	79.17	72.00	75.25	76.17	60.45
	KNN	85.12	85.12	83.94	87.96	87.14	89.78	84.78	86.12	87.95	72.78
	SVM	88.74	87.82	87.60	90.66	88.20	92.50	88.29	88.69	89.04	73.86
Entropy	DT	79.54	76.78	75.45	81.95	80.45	78.19	72.48	79.14	80.08	63.15
	KNN	87.44	85.43	84.75	88.97	91.78	88.74	85.46	88.47	85.45	76.18
	SVM	90.59	86.01	85.87	90.17	92.30	93.90	86.65	91.00	92.60	80.45
MTF	DT	80.15	78.24	81.17	82.84	79.48	80.16	82.78	82.82	75.48	73.15
	KNN	89.45	88.12	89.58	87.14	90.78	91.72	91.75	91.89	82.73	82.76
	SVM	92.05	91.12	90.60	89.92	93.21	93.29	92.72	93.05	93.21	84.62
TF-MTF	DT	80.48	79.16	79.18	82.95	80.49	80.79	81.73	83.49	75.94	73.46
	KNN	89.15	89.45	88.16	88.58	91.46	92.86	89.74	91.47	83.15	82.89
	SVM	92.41	90.20	89.59	90.40	93.11	93.50	90.80	92.66	93.76	85.14
TFIDF-MTF	DT	82.17	80.78	83.78	83.17	80.94	83.78	82.17	85.17	76.16	73.79
	KNN	91.16	89.47	91.89	90.47	91.76	92.45	93.08	92.16	83.93	83.74
	SVM	92.78	93.50	93.19	91.44	93.51	93.79	92.98	93.76	84.79	86.98
Entropy-MTF	DT	82.45	81.17	84.47	83.58	81.54	84.97	82.47	85.86	77.76	73.86
	KNN	92.28	90.26	88.16	91.72	92.47	91.47	93.16	92.18	84.45	85.48
	SVM	94.15	93.97	92.40	92.46	93.72	93.66	92.48	93.46	95.68	87.8

keyterms k_2, k_3 and k_4 commonly occur in the same documents where k_1 occurs. So calculate the number of occurrence of the keyterms k_1, k_2 and k_3 in the documents d1, d2, d4 and d5 as shown in Fig. 3. If no keyterms commonly occur with the keyterm k_i, then calculate the frequency based on the classical TF-IDF as shown in Table 3.

From Table 3, it is clearly identified that the remaining keyterms k1-4 do not occur commonly with the keyterm k5 in all the documents. So find only the number of occurrences of the keyterm k5 in all the documents for assigning weight to it. The proposed work is an extended method of classical TF-IDF formula. So the value of MTF is multiplied with the value of classical TF-IDF as shown in Eq. 9:

$$W_{dj} = TF - IDF_{t,d} \times MTF_{t,d} \quad (9)$$

Performance Measure: Precision, recall and F-score are the broadly used measures to assess the result in the field about IR (Liu and Peng, 2014; Sabbah and Selamat, 2014). The proposed method also evaluates the performance based on these three measures. Precision defines the ratio of the number of relevant documents retrieved and the total number of retrieved document. Recall characterizes the proportion of the amount of documents effectively retrieved and the downright amount for documents. F-score is measured by blending precision and recall used to evealuate the effectiveness of the proposed work. The three measurements precision, recall and f-score are described in Eq. 10-12, respectively. The following definitions are used in equation to define the corresponding definitions:

- True Positive (TP); Number of documents correctly identified by the classifier
- False Positive (FP); Number of documents incorrectly identified by the classifier
- False Negative (FN); Number of non documents incorrectly identified by the classifier

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \quad (10)$$

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|} \quad (11)$$

$$f - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} \times \text{Recall}} \quad (12)$$

Classification: In late years, real world databases are expanded quickly (Fayyad *et al*., 2003; Lyman and Varian, 2003). So, the requirement with extricate knowledge from very substantial databases are also expanding. Knowledge Discovery in Databases (KDD) (Fayyad *et al*., 1996) could be characterized like non-trivial procedure to identify valid, novel, possibly useful and eventually justifiable patterns on information. Data mining is the specific pattern recognization undertaking in the KDD procedure.

Web Page Classification (WPC) is an important and difficult task in datamining. It assits the clients to acquire majority of useful data from the enormous web which is more effective. WPC is a critical strategy for data analysis to classify the data classes. Data classification is a two step process. First, a model is fabricated to utilize a known

set in the data classes is called as training dataset. Second, the derived model tries to utilize testing data with the training samples. The accuracy of the model is measured by distinguishing the known model classes with the expected model classes. There are lot of procedures available for classification of data such as Decision Tree (DT), Bayesian networks, Fuzzy logic Neural Networks, K-Nearest Neighbor (KNN), Support Vector Machines (SVM) and so on. In Three classifiers namely SVM, K-NN, DT have been used in this study for classification.

Support Vector Machine (SVM): The SVM (Sabbah and Selamat, 2014; Sabbah *et al*., 2016) is used for learning classification and regression conditions from data which is used to learn radial basis function (RBF), polynomial and Multi-Layer Perceptron (MLP) classifiers. SVMs were proposed for classification by Vapnik in the 1960s. SVM is an efficient machine learning algorithm for data classification. This procedure will be used to attain the higher accuracy in the procedure about retrieval.

SVM has several advantages as text classifier (Sabbah *et al*., 2016). It can handle exponential features and infinitely features. Redundant features, high dimensional features are well handled by SVM and it does not require a destructive feature selection. SVM has proved to be one of the best and accurate classifier methods in many other domains. K-Nearest Neighbor (K-NN). It is a non-parametric strategy used for classification and regression (Khamar, 2013; Sabbah *et al*., 2016) based on distance and similarity function like Euclidean distance function. It may be attempted to huge numbers of application due to its effectiveness, non-parametric and not difficult to completing properties.

In K-NN, the input has k nearest training samples and the output depends on the usage of K-NN. The K-NN may be used for either classification or regression. In k-NN classification, the output, i.e., object is classified based on its neighbors with object being assigned to the class of that K nearest neighbors. In K-NN regression, the output is the average value of objects k nearest neighbors Decision Tree (DT). DT classifier will be an easy and generally utilized classification procedure (Vivekanandan and Karpagavalli, 2014) which applies simple idea to classification. DT classifier creates a set of questions regarding those attributes of the test record. Every chance it get an answer, a question is followed until a conclusion is arrived about the class label of the record. In DT, those root and internal nodes hold attribute test condition will differentiate records that bring different characteristics. The whole terminal nodes will be allocated a class label Yes or No.

When the DT is constructed, classifying a test record will be simple. Beginning from the root node, apply those test conditions to the record and based on the output of test, follow the branch. It will be followed to a leaf node or another internal node when the new test condition is appeared. If the leaf node is reached, the class label of that node will be assigned into the record.

RESULTS AND DISCUSSION

Based on the discussion in the previous section, experiments have been conducted on four different term weighting methods such as TF, TF-IDF, Entropy and MTF on 20 newsgroup dataset. After weight is assigned to the documents the different classifiers such as DT, KNN and SVM are applied. Table 3 gives the f-score values different classification methods applied on term weighting methods like TF, TF-IDF, Entropy and MTF.

Figures 3-5 show the f-score values of different classification methods such as DT, KNN and SVM. In this proposed work the MTF weighting method is also combined with other term weighting methods discussed as. From Fig. 3 it can be seen that the DT classification performance of the different term weighting schemes entropy and MTF outperform the other term weighting schemes such as TF, TF-IDF in f-score.

In addition, it can be seen that the f-score value of the proposed MTF value is higher in most of the categories when compared to the other term weighting methods. MTF achieves the maximum DT classification value of 82.84 in the category misc.forsale. Also from Fig. 3 it can be identified that the proposed method MTF combined with the existing term weighting methods has the better f-score values. The f-score values of KNN classifier for different term weighting methods is illustrated in Fig. 4. In that, f-score values of entropy is closely related the proposed work in most of the categories. However, MTF produces the better classification than entropy. In category sci.space, f-score value of MTF is 91.89 which is the highest f-score value but in the same category entropy is 88.47. The reason behind this is that the MTF consider the frequency of the co-terms commonly occur in a document.

Figure 5 shows the comparison view of SVM classification performance of different term weighting methods in terms of f-measure. The MTF classification performance is higher than the other term weighting methods such as TF, TF-IDF and Entropy. In addition, it is clearly identified that the classification performance of combined term weighting methods outperform the other methods. The highest value of f-score is achieved as 95.68 n the category soc.religion.christian. Figure 6-8 shows the different classification performance of the proposed MTF.

	k1	k2	k3	k4	k5
d1	35	24	0	23	17
d2	0	23	14	0	27
d3	28	17	0	25	20
d4	27	24	0	31	15
d5	0	0	15	0	14
d6	0	15	32	16	0

	k5
d1	17
d2	27
d3	20
d4	15
d5	14
∑	93

Document Weight	d1	17/93	0.1828
	d2	27/93	0.2903
	d3	20/93	0.2151
	d4	15/93	0.1613
	d5	14/93	0.1505

Fig. 3: Weight calculation based on the keyterm k1

	k1	k2	k3	k4	k5
d1	35	24	0	23	17
d2	0	23	14	0	27
d3	28	17	0	25	20
d4	27	24	0	31	15
d5	0	0	15	0	14
d6	0	15	32	16	0

	k1	k2	k4	k5	∑
d1	35	24	23	17	99
d3	28	17	25	20	90
d4	27	24	31	15	97
Total					286

Document Weight	d1	99/286	0.3462
	d3	90/286	0.3147
	d4	97/286	0.3392

Fig. 4: Weight calculation based on the keyterm k5

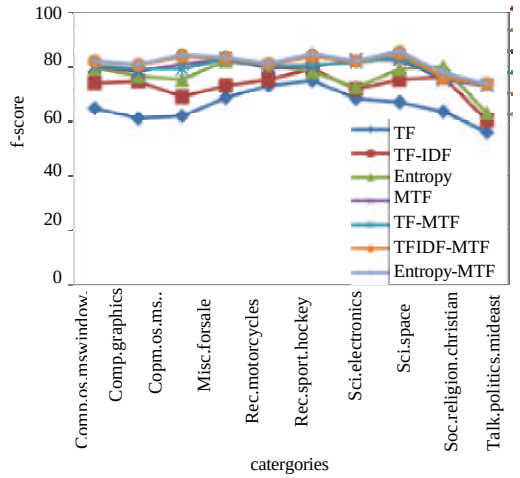

Fig. 5: f-score values for KNN classifier

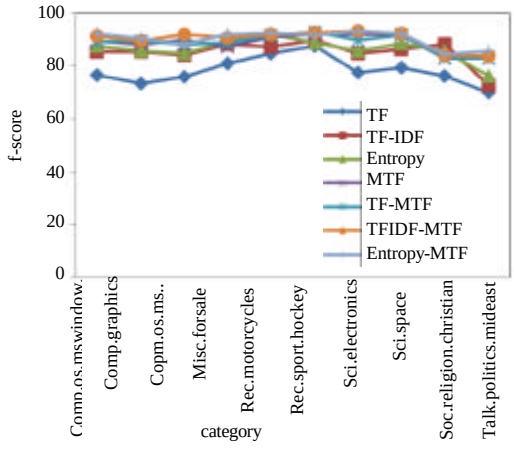

Fig. 6: F-score values for KNN classifier

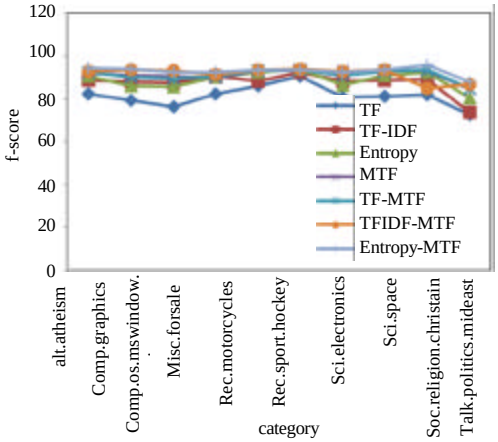

Fig. 7: F-score values for SVM classifier

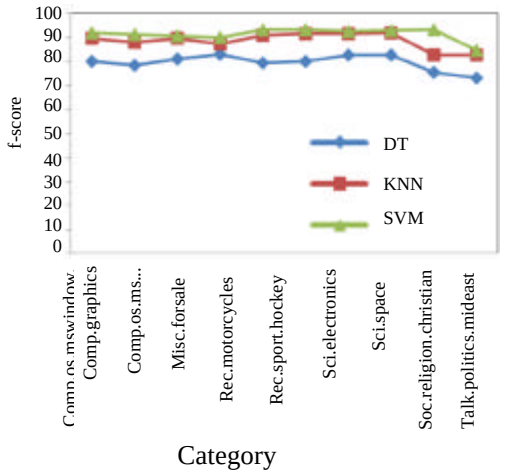

Fig. 8: Comparsion of classifiers for proposed MTF

CONCLUSION

This study proposes a modified frequency based term weighting method called MTF. It is the extended work of classical TF-IDF. It assigns weight to the term based on the occurrence and also the occurrence of the co-terms in same document. Based on the weight of a document the classification accuracy has been calculated by three different classifiers such as DT, KNN and SVM for the term weighting methods like TF, TF-IDF and MTF. In all these classifiers the proposed method MTF has the better classification accuracy than the other weighting methods. In this research, the proposed method has also been combined with existing term weighting methods. It also give the better classification accuracy than other weighting methods.

ACKNOWLEDGMENTS

The first research thanks to University Grants Commission (UGC) for providing financial support for his research work under the grant Rajiv Gandhi National Fellowship from the academic year 2013-14.

REFERENCES

Attia, M., L. Tounsi, P. Pecina, J.V. Genabith and A. Toral, 2010. Automatic extraction of Arabic multiword expressions. Proceedings of the the 7th Conference on Language Resources and Evaluation (LREC 2010), June 7, 2010, DORAS, Beijing, China, pp: 18-26.

Bouchekif, A. G. Damnati and D. Charlet, 2014. Intra-content term weighting for topic segmentation. Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 4-9, 2014, IEEE, Florence, Italy, pp: 7113-7117.

Carmel, D., A. Mejer, Y. Pinter and I. Szpektor, 2014. Improving term weighting for community question answering search using syntactic analysis. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, November 3-7, 2014, ACM, New York, USA., ISBN: 978-1-4503-2598-1, pp: 351-360.

Chiang, D.A., H.C. Keh, H.H. Huang and D. Chyr, 2008. The Chinese text categorization system with association rule and category priority. Expert Syst. Appl., 35: 102-110.

Choi, D., B. Ko, H. Kim and P. Kim, 2014. Text analysis for detecting terrorism-related articles on the web. J. Netw. Appl., 38: 16-21.

Doko, A., M. Stula and D. Stipanicev, 2013. A recursive TF-ISF based sentence retrieval method with local context. Int. J. Mach. Learn. Comput., 3: 195-200.

Fang, H., T. Tao and C. Zhai, 2011. Diagnostic evaluation of information retrieval models. ACM. Trans. Inf. Syst., 29: 1-49.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996. From data mining to knowledge discovery in databases. AI Mag., 17: 37-54.

Fayyad, U.M., Shapiro, G.P. and R. Uthurusamy, 2003. Summary from the KDD-03 panel: Data mining: The next 10 years. ACM. SIGKDD. Explorations Newsl., 5: 191-196.

Gautam, J. and E. Kumar, 2013. An integrated and improved approach to terms weighting in text classification. IJCSI. Int. J. Comput. Sci. Issues, 10: 310-314.

Goswami, P. and V. Kamath, 2014. The DF-ICF algorithm-modified TF-IDF. Int. J. Comput. Appl., 93: 28-30.

Huo, W., 2012. Automatic multi-word term extraction and its application to web-page summarization. Ph.D. Thesis, The University of Guelph, Guelph, Ontario, Canada.

Keh, H.C., D.A. Chiang, C.C. Hsu and H.H. Huang, 2010. The chinese text categorization system with category priorities. J. Software, 5: 1137-1143.

Khamar, K., 2013. Short text classification using kNN based on distance function. IJARCCE. Int. J. Adv. Res. Comput. Commun. Eng., 2: 1916-1919.

Lan, M., C.L. Tan and H.B. Low, 2006. Proposing a new term weighting scheme for text categorization. Proceedings of the 21st National Conference on Artificial Intelligence, pp: 763-768.

Lan, M., C.L. Tan, J. Su and Y. Lu, 2009. Supervised and traditional term weighting methods for automatic text categorization. Pattern Anal. Mach. Intell. IEEE. Trans., 31: 721-735.

Lee, D.L., H. Chuang and K. Seamons, 1997. Document ranking and the vector-space model. Software IEEE., 14: 67-75.

Liangtu, S. and Z. Xiaoming, 2007. Web text feature extraction with particle swarm optimization. IJCSNS. Int. J. Comput. Sci. Netw. Secur., 7: 132-136.

Liu, L. and T. Peng, 2014. Clustering-based method for positive and unlabeled text categorization enhanced by improved TFIDF. J. Inf. Sci. Eng., 30: 1463-1481.

Liu, Y., Y. Wang, L. Feng and X. Zhu, 2014. Term frequency combined hybrid feature selection method for spam filtering. Pattern Anal. Appl., 1: 1-15.

Lyman, P. and H.R. Varian, 2003. How much information 2003? http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

Murray, G. and S. Renals, 2007. Term-Weighting for Summarization of Multi-Party Spoken Dialogues. In: Machine Learning for Multimodal Interaction. Popescu, B.A., S. Renals, B. Herve (Eds.). Springer Berlin Heidelberg, Berlin, Germany, pp: 156-167.

Paik, J.H., 2013. A novel TF-IDF weighting scheme for effective ranking. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28-August 01, 2013, ACM, New York, USA., ISBN: 978-1-4503-2034-4, pp: 343-352.

Pesaranghader, A., N. Mustapha and N.M. Sharef, 2013. Improving multi-term topics focused crawling by introducing Term Frequency-Information Content (TF-IC) measure. Proceedings of the 2013 International Conference on Research and Innovation in Information Systems (ICRIIS), November 27-28, 2013, IEEE, Kuala Lumpur, Malaysia, ISBN: 978-1-4799-2486-8, pp: 102-106.

Porter, M.F., 1980. An algorithm for suffix stripping. Program Electron. Lib. Inform. Syst., 14: 130-137.

Raj, R.G., 2012. Improving the relevancy of document search using the multi-term adjacency keyword-order model. Malaysian J. Comput. Sci., 25: 1-10.

Ricardo, R.M.B., 2013. Ranking of multi-word terms. Master Thesis, Leiden Institute of Advanced Computer Science, Leiden University, Netherlands

Sabbah, T. and A. Selamat, 2014. Modified Frequency-Based Term Weighting Scheme for Accurate Dark Web Content Classification. In: Information Retrieval Technology. Jaafar, A., N.M. Ali, , S.A.M. Noah, A.F. Smeaton and P. Bruza et al. (Eds.). Springer International Publishing, New York, USA., pp: 184-196.

Sabbah, T., A. Selamat, M.H. Selamat, R. Ibrahim and H. Fujita, 2016. Hybridized term-weighting method for dark web classification. Neurocomput., 173: 1908-1926.

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, Boston, MA., USA., ISBN: 9780201122275, Pages: 530.

Sanderson, M. and I. Ruthven, 1996. Report on the glasgow IR group (glair4) submission. Proceedings of the Fifth Text Retrieval Conference (TREC-5), November 20-22, 1996, Gaithersburg, Maryland, USA., pp: 517-520.

Santhanakumar, M. and C.C. Columbus, 2015a. Web usage based analysis of web pages using RapidMiner. WSEAS. Trans. Comput., 14: 455-464.

Santhanakumar M. and C.C. Columbus, 2015b. Various improved TFIDF schemes for term weighting in text categorization: A survey. Int. J. Appl. Eng. Res., 10: 11905-11910.

Selamat, A. and S. Omatu, 2003. Neural networks for web page classification based on augmented PCA. Proceedings of the International Joint Conference on Neural Networks 2003, July, 20-24, 2003, IEEE, New Jersey, USA., pp: 1792-1797.

Vivekanandan, M.V. and M.N. Karpagavalli, 2014. Efficient data analysis scheme for increasing performance in big data. Int. J. Res. Sci. Technol., 1: 193-198.

Wang D. and H. Zhang, 2013. Inverse-category-frequency based supervised term weighting scheme for text categorization. J. Inf. Sci. Eng., 29: 209-225.

Wang, N., P. Wang and B. Zhang, 2010. An improved TF-IDF weights function based on information theory. Proceedings of the 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering (CCTAE), June 12-13, 2010, IEEE, Chengdu, China, pp: 439-441.

Wang, Y., Y. Liu, L. Feng and X. Zhu, 2015. Novel feature selection method based on harmony search for email classification. Knowledge Based Syst., 73: 311-323.

Xia, T. and Y. Chai, 2011. An improvement to TF-IDF: Term distribution based term weight algorithm. J. Software, 6: 413-420.

Zaefarian, R., J. Siddiqi, B. Akhgar and G. Zaefarian, 2006. A new algorithm for term weighting in text summarization process. Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, August 18-20, 2006, World Scientific and Engineering Academy and Society (WSEAS), Corfu Island, Greece., pp: 292-297.

Zulkifeli, W.W., N. Mustapha and A. Mustapha, 2012. Classic term weighting technique for mining web content outliers. Int. Conferenceon Comput. Tech. Artif., 1: 271-275.