An Efficient Term Weighting Approach For Document Classification using Knn Classifier

Aijazahamed Qazi, R. H. Goudar, P.S.Hiremath

Abstract*: With substantial expansion in the volume of computerized information, document classification has become an emerging area of exploration in the research community. A defined methodology for this task is to apply machine learning methods. The traditional term frequency and inverse document frequency weighting approach considers only the statistical information. This paper introduces a new approach to weight a term by calculating the semantic similarity between the category label and the term. Also, the weight of a term comprises of its co-occurrence computation. Experiments were carried on the Reuters-21578 benchmark dataset. The results obtained specify that the proposed method outperforms the traditional method with kNN classifier.

***Index Terms*: Term frequency, kNN, Similarity, Ontology.**

I. INTRODUCTION

The growth in the amount of electronic documents has made information retrieval more thought-provoking task to discover significant data. As web documents have metadata, classification of documents is becoming more challenging as discussed by *Mironczuk et al.* in [1]. A feature vector represents the document in the vector space model and is assigned to the categories which are predefined using machine learning methods. The efficiency of an information retrieval model depends on the technique that is applied to weight a term. Term frequency-Inverse document frequency(TF-IDF) is an extensively used weighting method in field of information retrieval. TF-IDF does not consider the semantic information related to the term. Weighting decides the prominence of a term in a document as discussed by *Qazi et al.* in [2]. As an information retrieval model deals with imprecise knowledge, evaluation of user's query by web search engine is one of the search problems. The following are the key contributions of this paper: Initially, the proposed term weighting approach captures the semantic similarity between the categories and the terms. Secondly, a comparative analysis on TF-IDF and proposed term weighting is carried by conducting experiments on Reuters-21578 collection. The paper is structured as follows: Section 2 reviews various term weighting methods. Section 3 provides an overview of the preliminaries required. Section 4 elaborates the proposed term weighting approach. The experimental procedure and result analysis is presented in Section 5. Section 6 draws the conclusion of the paper.

II. RELATED WORK

An information retrieval framework begins with the lexical analysis of documents and leads to retrieval where user queries are matched against documents. Term weight is the statistical information that specifies the significance of a term in the document. Supervised and Unsupervised approaches broadly assign weights to the terms in a document collection. Any method to weight a term attempts to improve the classification accuracy of document. TF-IDF weights a term, relative to its occurrence in a given document, and inversely proportionate to its occurrence in the document collection. Sebastiani in [3] surveyed and discussed the essential approaches to text categorization using machine learning, document representation and performance evaluation. Debole et al. in [4] proposed a supervised term weighting methodology by applying category based term assessment function. Peng et al. in [5] presented an improved TF-IDF weighting approach with a voting classifier for better results. Chen et al. in [6] proposed a Term Frequency-Inverse Gravity Moment based weighting method to calculate the influence of a term within the class. The efficiency was evaluated with SVM and kNN classifiers on benchmark datasets. Jiang et al. in [7] proposed a correlation-based weight assignment. In correlation-based feature weighting, the weight was assigned to a feature with sigmoid variation. Feng et al. in [8] proposed a technique to weight a term with probabilistic model. Jiang et al. in [9] presented a term weighting approach by applying the kNN classifier to Reuters-21578 dataset. Al-anzi et al. in [10] investigated the performance of the text classification for arabic language. The feature extraction was based on Latent Semantic Indexing. Ma et al. in [11] introduced an ontology-based methodology to summarize and compute the similarity for classifying chinese judgment documents. Further, experiments were carried with kNN classifier and achieved higher classification accuracy. Elhadad et al. in [12] presented a technique based on the ontological structure of the WordNet, to remove non-semantic words having no relation with any of the WordNet groups. As a result, the obtained feature vector was enriched by concatenating each word with its analogous WordNet group. Experiments were evaluated for classifiers like J48, Naive-Bayes, SVM and kNN. Qazi et al. in [13] proposed an ontology-based term weighting technique based on a domain to select the features. Results obtained showed significant improvement in the prediction performance.

Revised Manuscript Received on June 05, 2019

Aijazahamed Qazi, Department of CSE, SDMCET, Dharwad, India.
R.H.Goudar, Department of CNE, Center for PG Studies Visvesvaraya Technological University, Belgaum, India.
P.S.Hiremath, Department of MCA, KLE Technological University, Hubli, India.

Figueiredo et al. in [14] proposed an approach to create distinctive compound features, collected by the terms co-occurring in the documents..

This approach was tested with a few classifiers and improved accuracy was obtained. *Zhang et al.* in [15] proposed a cross relation analysis approach by combining co-occurrence and semantic associations for topic discovery. The proposed approach incorporated several associations into a graph and identified topics. *Lin et al.* in [16] proposed an approach to calculate the similarity amongst two clusters of documents.
WordNet is a semantic lexical repository designed by Princeton University. In the WordNet, verbs and nouns representing a word are organized into groups called synsets. WordNet offers minor descriptions and semantic relations between these alternative word sets was discussed in *Luo et al.* in [17]. *Li et al*. in [18] proposed a novel scheme to weight a term and exploit the semantics of terms and classes. The semantics of classes were characterized by the understanding of WordNet with the weight of a term associated to its category. Experimental results obtained predicted that the proposed method outperformed TF-IDF with a small amount of training data. *Miller* in [19] proposed an approach based on corpus-based vocabulary and WordNet to increase the performance of text categorization. Back propagation neural network and kNN algorithms were applied for experimentation. *Trstenjak et al.* in [20] proposed a framework for text cataloging using kNN algorithm. Further, an investigation was conducted and kNN model based classifier was suggested. An experimental evaluation was carried out on the 20-newsgroups and Reuters-21578 collection.
Therefore, the purpose of our study is to compare the TF-IDF weighting method with our proposed method on Reuters-21578 dataset with kNN classifier.

III. PRELIMINARIES

A. Term Frequency-Inverse Document Frequency

Traditional approaches to weight a term are Boolean, Term Frequency and TF-IDF. TF-IDF is a weighting procedure used in the field of information retrieval. TF-IDF is a statistical quantity used to estimate the influence of a word in the collection [3]. Term Frequency refers to the count of term's occurrence in a document. The significance of a term surges to the quantity that a word appears in the document. Inverse Document Frequency measures the significance of a term in a collection. Variants of TF-IDF are applied by search engines to assess the relevance of a user's query the document. TF-IDF weight for a term t_k in a document is indicated as follows,

$$w(t_k) = tf_k . log\left(\frac{N}{df_k}\right) \quad (1)$$

B. Lin's Semantic Similarity

The interpretation of textual information involves the evaluation of semantic likeness amongst the terms. Ontology provides description of concepts and relationships that exist between them. The concept of Information Content refers to the probability of occurrence of each concept in the corpus.

$$IC(a) = -log\ P(a) \quad (2)$$

Semantic similarity can be determined by the common information shared amongst the two terms which are characterized by their Least Common Subsumer (LCS) in the ontology.

$$sim_{lin}(a,b) = \frac{2xIC(LCS(a,b))}{(IC(a) + IC(b))} \quad (3)$$

C. kNN Algorithm

The k-nearest-neighbor is a lazy learner and non-parametric machine learning algorithm. kNN classifier has been largely applied in the field of pattern recognition. kNN classifier is based on learning by correlating a test vector with a training vector. For a test vector to be classified, kNN classification algorithm examines the k trained vectors neighboring to the test vector. The test vector is given the label of k nearest neighbors. The Euclidean distance calculates the proximity between the two vectors. The Euclidean distance amongst two vectors, X = (x_1, x_2, … , x_j) and Y = (y_1, y_2, …, y_j) is

$$d_{x,y} = \sqrt{\sum_{j=1}^{J}(x_j - y_j)^2} \quad (4)$$

IV. PROPOSED TERM WEIGHTING APPROACH

Algorithm *: Computing Document Term Weight Matrix*
Input: *Corpus D of Web Documents*
Output*: $W \leftarrow W_{ij}$, Document Term Weight Matrix of size $|D|$ x $|T|$*
Procedure:
for each $d_i \in D$ do
Remove stopwords and construct set T of tokens in d_i
for each term $t_i \in T$ do
Compute term frequency tf_i
end for
end for
Construct the co-occurrence matrix C_T of size $|T|$ x $|T|$ with ij^{th} element $C_T(i,j)$ being the number of joint occurrences of i^{th} and j^{th} terms in a document of D
for each i^{th} row of C_T do
Compute rowsum $\alpha_i \leftarrow \sum_{j=1}^{|N|} C_T(i,j) / |T|$
end for
Let $CL=\{l_1, l_2, \ldots, l_8\}$ be the set of labels of 8 categories of web documents in D
for each label $l_k \in CL$ do
Extract the first synonym S_{l_k} *from the wordnet*
end for
Construct the set of synonym labels,
$S_{C_L} = \{S_{L_1}, S_{L_2}, \ldots, S_{L_8}\}$
for each term $t_i \in$ do
Compute $\beta i \leftarrow \max_{s_{l_k} \in S_{CL}}(SIM_{LIN}(t_i, S_{l_k}))$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

where ,

$SIM_{LIN}(x,y) \leftarrow 2\log(LCS(x,y))/(\log(P(x)) + \log(P(y)))$

with x,y ← terms,
LCS(x,y) ← least common subsumer of terms
end for
Compute modified term frequency mtf_i for each $t_i \in T$,
$mtf_i \leftarrow tf_i + \alpha_i + \beta_i$
for each document $d_j \in D$ do
for each term $t_i \in T$ of dj do
Compute, $W_{ij} \leftarrow mtf_i$ x log($|D|/df_i$)
where $|D| \leftarrow$ total number of documents
$df_i \leftarrow$ number of documents containing term ti
end for
end for
Normalize each row vector of W

The algorithm describes the proposed term weighting technique for document classification. Initially textual content is preprocessed. For the features extracted, modified term frequency is computed for each document. Then the inverse document frequency is calculated.

V. EXPERIMENTS

A. Dataset

Reuters-21578 is one of the benchmark dataset used for document classification. This collection has 21,578 news articles with a set of 118 subject categories. A split of the collection, called ModApte split, which includes 9,603 training and 3,299 test documents is considered. Further, the dataset has 90 categories with a training set of 7769 documents and a test set of 3019 documents. The experiments were conducted on eight categories of the Reuters-21578 collection.

B. Results and Analysis

The training and test documents were subjected to pre-processing, which comprised of removal of stopwords, stemming, feature selection and feature weighting. Experimental results obtained are based on the F1-Measure. The experiment considers the train-test split. The value of k for kNN algorithm is varied and the scores obtained with TF-IDF and proposed approach is presented in Table 1. 80% of the test documents were classified accurately with the proposed term weighting approach.

Category	TF-IDF	Proposed approach
Acq	0.67	0.80
Corn	0.71	0.79
Crude	0.65	0.74
Earn	0.68	0.82
Interest	0.70	0.81
Ship	0.71	0.77
Trade	0.69	0.79
Oilseed	0.68	0.77
avg.Micro-F1	0.69	0.81
avg.Macro-F1	0.65	0.76

Table 1. kNN performance on Reuters-21578 subset

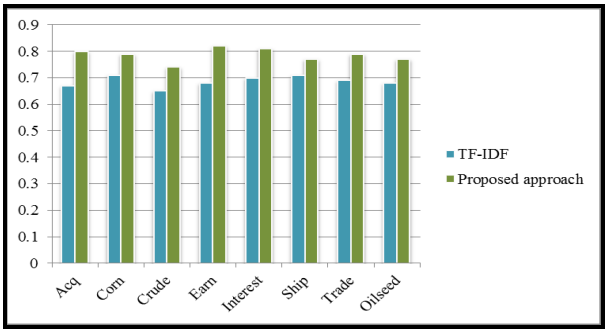

Fig 1: Performance analysis

VI. CONCLUSION

The paper provides an overview of a novel weighting scheme for document classification. This paper provides an experimental assessment between two-term weighting methods and infers that proposed method yields better accuracy for document classification. As a part of future scope, we propose to investigate and apply the technique for multi-label document classification.

REFERENCES

1. Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications.
2. Qazi, A., & Goudar, R. H. (2016). Emerging Trends in Reducing Semantic Gap towards Multi-media Access: A Comprehensive Survey. Indian Journal of Science and Technology, 9, 30.
3. Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1 (March 2002), 1-47.
4. S Franca Debole and Fabrizio Sebastiani. 2003. Supervised term weighting for automated text categorization. In Proceedings of the 2003 ACM symposium on Applied computing (SAC '03). ACM, New York, NY, USA, 784-788
5. Tao Peng, Lu Liu, and Wanli Zuo. 2014. PU text classification enhanced by term frequency-inverse document frequency-improved weighting. Concurr. Comput. : Pract. Exper. 26, 3 (March 2014), 728-741.
6. Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Systems with Applications, 66, 245-260.
7. L. Jiang, L. Zhang, C. Li and J. Wu, "A Correlation-Based Feature Weighting Filter for Naive Bayes," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 2, pp. 201-213, 1 Feb. 2019.
8. Feng, G., Li, S., Sun, T., & Zhang, B. (2018). A probabilistic model derived term weighting scheme for text classification. Pattern Recognition Letters, 110, 23-29.
9. Jiang, H., Li, P., Hu, X., & Wang, S. (2009, November). An improved method of term weighting for text classification. In 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems (Vol. 1, pp. 294-298). IEEE.
10. Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Journal of King Saud University-Computer and Information Sciences, 29(2), 189-195.
11. Ma, Y., Zhang, P., & Ma, J. (2018). An Ontology Driven Knowledge Block Summarization Approach for Chinese Judgment Document Classification. IEEE Access, 6, 71327-71338.
12. Elhadad, M. K., Badran, K. M., & Salama, G. I. (2018). A Novel Approach for Ontology-Based Feature Vector Generation for Web Text Document Classification. International Journal of Software Innovation (IJSI), 6(1), 1-10.
13. Qazi, A., & Goudar, R. H. (2018). An Ontology-based Term Weighting Technique for Web Document Categorization.Procedia computer science, 133, 75-81.
14. Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira Jr, W. (2011). Word co-occurrence features for text classification. Information Systems, 36(5), 843-858.

15. Zhang, C., Wang, H., Cao, L., Wang, W., & Xu, F. (2016). A hybrid term–term relations analysis approach for topic detection. Knowledge-Based Systems, 93, 109-120.
16. Lin, Y. S., Jiang, J. Y., & Lee, S. J. (2014). A similarity measure for text classification and clustering. IEEE transactions on knowledge and data engineering, 26(7), 1575-1590.
17. Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. Expert Systems with Applications, 38(10), 12708-12716.
18. Li, C. H., Yang, J. C., & Park, S. C. (2012). Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet. Expert Systems with Applications, 39(1), 765-772.
19. Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.
20. Trstenjak, B., Mikac, S., & Donko, D. (2014). kNN with TF-IDF based Framework for Text Categorization. Procedia Engineering, 69, 1356-1364.

AUTHORS PROFILE

Aijazahamed Qazi, currently working as an Assistant Professor, Dept. of CSE, SDMCET, Dharwad. He has published papers in International Journals and Conferences. His areas of interest include Semantic Web and Information Retrieval.

Dr.R.H.Goudar, currently working as an Associate Professor, Dept. of CNE, Visvesvaraya Technological University, Belagavi. He has 14 years of teaching experience at Professional Institutes across India. He worked as a faculty at International Institute of Information Technology, Pune for 4 years and at Indian National Satellite Master Control Facility, Hassan, India. He has published over 130 papers in International Journals, Book Chapters and Conferences of high repute. His subjects of interest include Semantic Web, Network Security and Wireless Sensor Networks.

Dr.P.S.Hiremath, currently working as a Professor, Dept. of MCA, KLE University, Hubli. He has published several papers in International Journals and Conferences. He has guided many research scholars. His areas of interest include Image Processing, Pattern Recognition and Natural Language Processing.