ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 5(J), pp. 27191-27194, May, 2018

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

A NOVEL APPROACH FOR WORD NET ONTOLOGY BASED DOCUMENT CATEGORIZATION TECHNIQUE

Sukhmani and Chauhan R.K

Department of Computer Science and Applications, Kurukshetra University

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0905.2207

ARTICLE INFO

Article History:

Received 08th February, 2018
Received in revised form 12th March, 2018
Accepted 26th April, 2018
Published online 28th May, 2018

Key Words:

Document classification, wordnet, ontology, semantic, vector space model.

ABSTRACT

The assignment of mining, the utilized component vector is estimated as the most critical errand for enhancing the content handling abilities. In this paper, authors have proposed another approach which is to be utilized as a part of the arrangement of highlight vector for report grouping of web content. Semantics are appended with highlight vector. Utilization of wordnet ontology helps in expulsion of futile word. This outcomes in enhanced component vector estimate as size of highlight vector has been lessened because of the expulsion of futile words. Content reports are spoken to through vector space display. Proposed wordnet ontology based strategy was contrasted and some non wordnet philosophy based methods and it has been discovered that proposed system has beated the other on all the pertinent parameters.

INTRODUCTION

Document Categorization (DC) or Text Categorization (TC) is the grouping of archives as for an arrangement of at least one previous classes. Record Categorization is a hard and exceptionally helpful activity regularly connected to dole out subject classifications to reports, to course and channel writings, or as a piece of characteristic dialect handling system[1].

WordNet is a lexical database for the English language[2]. It organizes English words into units of equivalent words alluded to as synsets, bears fast definitions and use cases, and records some of relatives among those equivalent word units or their individuals. WordNet can hence be viewed as a total of lexicon and glossary. While its miles available to human clients through a web program, its essential utilize is in robotized content investigation and manufactured knowledge programs.

Synsets are interlinked by methods for theoretical semantic and lexical relatives. The subsequent group of definitively related expressions and thoughts might be explored with the program. WordNet is likewise unreservedly and openly accessible for download. WordNet's shape makes it a valuable gadget for computational phonetics and characteristic dialect processing[3].

Natural Language Processing (NLP) is a territory of software engineering and counterfeit consciousness worried about the communications amongst PCs and human (regular) dialects, specifically how to program PCs to productively process a lot of common dialect information.

The WordNet philosophy method had been utilized to lessen the dimensionality of the component vector for the web content grouping report.

This paper contains the five segments. The area I comprise of the presentation about the point of the paper. The segment II comprise of the related work or the past work. The area III comprise of the means required amid the period of the fulfillment. The segment IV comprise of the trial comes about subterranean insect the execution assessment. The last and the essential area V comprise of the conclusion.

Related Works

In this stage, we can quickly review a portion of the looks into executed. Mohamed K. Elhadad *et al*. (2017) proposed a particular strategy for cosmology based absolutely for web content document class. In this strategy he utilized the Vector Space Model (VSM) for mining obligations to speak to literary substance records and the Term Frequency Inverse Document Frequency (TFIDF) is utilized as a day and age weighting procedure. The proposed philosophy basically based approach

Corresponding author:* **Sukhmani
Department of Computer Science and Applications, Kurukshetra University

changed into assessed towards the Principal Component Analysis (PCA) technique utilizing a few experiments[5].

Taeho Jo *et al*. (2017) proposed a string vector based KNN for record arrangement. The creator had proposed the string vector based variant of the KNN as the way to deal with the content classification. . Keeping in mind the end goal to take care of the issues, in this exploration, writings are encoded into string vectors, rather than numerical vectors, the comparability measure between string vectors is defined, and the KNN is modified into the adaptation where string vector is given its input[6].

Sujata R.Kolhe *et al*. (2017) proposed an idea driven report bunching approach utilizing WordNet. As of late Information Technology is utilized broadly for extensive variety of use for instance arrangements empowered through web based business to various online data framework. This utilization has prompt improvement of huge printed information base. For the most part this data information is put away in unstructured text[7].

Dashen Xue *et al*. (2015) proposed research of content order display in view of arbitrary woods. Because of the great execution in calculation speed and productivity, Random Forest (RF) calculation as a well known coordinated learning calculation has been broadly connected in numerous fields[9].

This paper proposes a way to deal with accomplish the arrangement of the web content records with the utilization of wordnet philosophy. WordNet metaphysics helps in evacuation of futile word. This outcomes in enhanced component vector measure as size of highlight vector has been diminished because of the evacuation of futile words. The execution of the grouping result has assessed with the utilization of the F-Measure, Precision, Recall and Classification exactness.

Ontology (Metaphysics) Based Classification Phases

Content Mining or Text Mining on a substantial gathering of records is normally a perplexing procedure. The most widely recognized approach to represent the archives is a pack of words(BOW) which considers the quantity of events of each term(word/expression) however overlooks the request.

Text Preprocessing: Preprocessing is one of the principle content mining calculation. Uysal *et al*. [10] have explored the effect of preprocessing assignments especially in the territory of content grouping. The preprocessing assignment as a rule comprise of errands, for example, tokenization, separating, lemmatization and stemming.

Tokenization: It is the errand of separating a character succession into pieces called tokens, and maybe in the meantime discard certain characters, for example, accentuation marks. The rundown of tokens is then used to advance processing [11].

Filtering or Separating: Filtering is typically done on records to evacuate a portion of the words. A typical sifting is stop-words expulsion. Stop words are the words often show up in the content without having much substance data [12].

Lemmatization: It is the errand that thinks about the morphological examination of the words, that is gathering together the different arched types of word so they can be dissected as a solitary thing. The lemmatization techniques attempt to outline structures to interminable tense and things to single frame. Keeping in mind the end goal to lemmatize the reports we should determine the parts of speech (POS) of each expression of the archives and in light of the fact that POS is terrible and blunder inclined.

Stemming: Stemming technique go for acquiring stem (root) of the inferred words. Stemming calculation are without a doubt dialect subordinate [13].

Highlighting Extraction Pseudo Code

```
/* Preprocessing */

for each record
        {
              do content sifting;
              see the record's dialect;
              apply stemming;
              check hinder words;
        }
```

Highlight Selection: Before the component choice stage, the element extraction stage happens. The component extraction stage points in preprocessing the information archives and after that separating the sack of words(BOW) that speak to these records. The component choice stage points in lessening the dimensionality of the removed sack of words(BOW), which depends on the chain of importance of the WordNet cosmology to dispense with words that has no connection with any of the WordNet lexical classifications. This element is utilized to get the ideal component determination. This can be accomplished by applying the WuPalmer Similarity. The WuPalmer Similarity is characterized as:

$$simwp(c1.c2)=2\times N \div N1+N2+2\times N$$

Here, the simwp indicates the similitude esteem. C1 and c2 are ideas. N is the commom parent idea. N1 and N2 are the situation of the ideas.

Vector Space Model: A vector is a point in a vector space and has length (from the root to the point) and heading. A vector space is characterized by an arrangement of directly free premise vectors. The premise vectors relate to the measurements or bearings of the vector space. Vector space demonstrate is a model for speaking to the content archives. It is utilized as a part of data sifting, data recovery and ordering. In the wake of applying the vector space display, an ideal vector space demonstrate is gotten.

Classifiers: Classifier is a machine learning apparatus where the objective trait is ostensible. Classifiers are utilized for the examination of the trial results to be done. In this paper four classifiers have been utilized. They are the K Nearest Neighbor, Naïve Bayes, Random Forest and the Support Vector Machine. These classifiers are utilized for the better outcomes.

RESULTS AND DISCUSSION

In this segment, we will talk about our exploratory setup and the outcomes for assessing the execution of our proposed work. Right off the bat we will discuss the dataset utilized. In this approach, Reuters-21578 informational index have been

utilized. The explanation behind utilizing Reuters-21578 dataset is that it utilizes sensable words than other dataset and furthermore sifted dataset is utilized.

Reuters-21578 Dataset: The Reuters-21578 dataset is the dataset which is utilized as a part of the content or archive arrangement tests. The information in the Reuters-21578 dataset was gathered by the Carnegie aggregate from the Reuters newsgroup in 1987.The dataset utilized is a separated dataset and the words utilized as a part of the dataset are the sensible words than the other datasets. The reuters dataset can be downloaded from the accompanying connection https://martin-thoma.com/nlp-reuters.

Table I Reuters Dataset Table

CATEGORY	#TRAINING	#TEST	TOTAL
Gold	79	20	99
Money-supply	145	16	161
Gnp	70	13	83
Cpi	70	9	79
Cocoa	50	13	63
Alum	39	11	50
Grain	40	11	51
Copper	44	10	54
Jobs	51	4	55
Reserves	45	8	53
Rubber	32	9	41
Iron-steel	30	17	47
TOTAL	695	141	836

This table demonstrates the records of the things utilized as a part of the dataset.

Assessment Criteria

The outcomes can be figured by utilizing the equations of exactness, review and the precision. Trial comes about detailed in this segment depend on the supposed "F1 measure", which is the symphonious mean of exactness and review.

$$\mathbf{F1(recall,\ precision)} = \frac{2\times \mathbf{recall}\times \mathbf{precision}}{\mathbf{recall}+\mathbf{precision}}$$

$$\mathbf{Precision} = \boldsymbol{true\ positive} \div (\boldsymbol{true\ positive} + \boldsymbol{true\ negative}) \times \mathbf{100}$$

$$\mathbf{Recall} = \boldsymbol{true\ positive} \div (\boldsymbol{true\ positive} + \boldsymbol{false\ negative}) \times \mathbf{100}$$

Exploratory Results

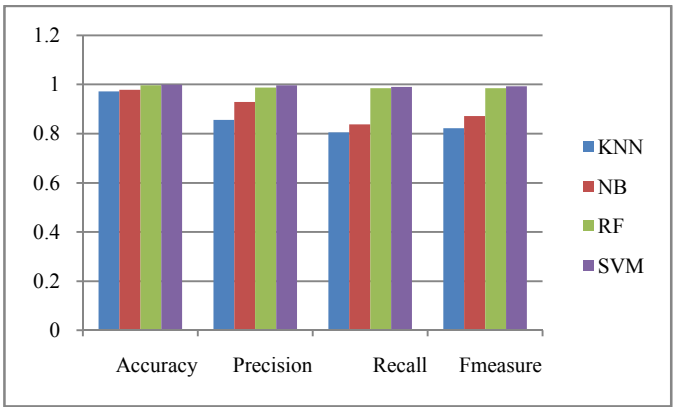
To test the approach for the WordNet Ontology based order system, different kinds of classifiers have been utilized. This approach has been performed on these four classifiers. These classifiers are K Nearest Neighbor, Naïve Bayes, Random Forest and Support Vector Machine (SVM).

Table II Assessment Measures Using Four Classifiers (KNN, NB, RF, SVM)

Approach	Accuracy	Precision	Recall	Fmeasure
KNN	0.972089	0.856364	0.80576	0.822409
NB	0.978868	0.9295	0.838269	0.871668
RF	0.99701	0.987939	0.984438	0.985552
SVM	0.998804	0.997006	0.98988	0.993311

In this table, the resultant execution utilizing four parameters (Fmeasure, Recall, Precision, Accuracy) of the web content archive have been demonstrated utilizing the four classifiers. As delineated in the table, the precision of the SVM classifier is more than alternate classifiers. SVM played out the best among every one of the classifiers utilizing the four parameters.

Chart Showing Experimental Result through Classifiers

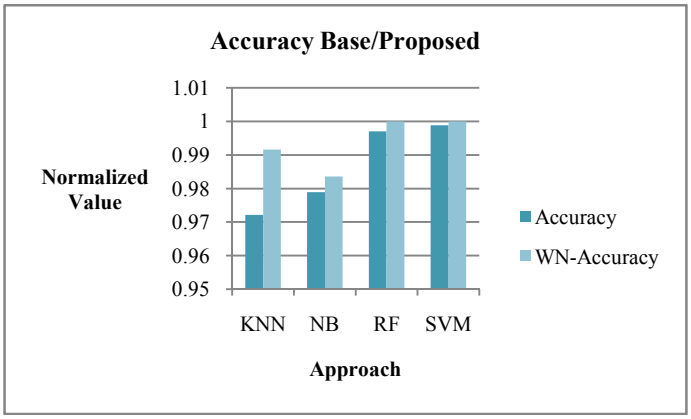


This chart additionally demonstrates that the exactness of the SVM classifier is more than alternate classifiers. SVM played out the best among every one of the classifiers utilizing the four parameters. After the SVM, the Random Forest is the best among alternate classifiers uniquely on account of the precision.

CONCLUSION

In this paper, a novel approach for the archive order is utilized. This approach depends on the WordNet Ontology. The WordNet Ontology is utilized as a part of the reference of the Term Frequency and the Inverse Document Frequency(TF-IDF) procedure. This approach is utilized to gauge the semantic comparability between the archives. In the wake of utilizing this system the outcomes are progressed.

Proposed Result



This diagram demonstrates that in the wake of applying the WordNet Ontology strategy, the exactness of the method has accomplished the better outcomes in contrast with alternate strategies.

References

1. Sebastiani F., “Machine Learning in Automated Text Categorization,” *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, (2002).
2. G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. (1990). WordNet: An online lexical database. *Int. J. Lexicograph*. 3, 4, pp. 235–244.
3. C. C. S. M.,. P.-J. C. Samia Iltache, "Using Domain Ontologies for Classification and Semantic Intepretation of Documents," in The Second International Conference

on Big Data, Small Data, Linked Data and Open Data, Portugal, (2016).
4. Elizabeth D Liddy.2001. Natural language processing. (2001)
5. Mohamed k. Elhabad "A Novel Approach for ontology-based dimensionality reduction for web text document classification " in ICIS (2017), Wuhan China
6. Taeho Jo "String Vector based KNN for Text Categorization" in ICACT (2017)
7. Mrs. Sujata R. Kolhe "A Concept Driven Document Clustering Using WordNet "in International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017)
8. Thamarai Selvi. S "Text Categorization using Rocchio Algorithm and Random Forest Algorithm" in IEEE eight International Conference on Advanced Computing (2016)
9. Dashen Xue "Research of text Categorization Model Based on Random Forests" in IEEE International Conference on Computational Intelligence and Communication Technology (2015)
10. Alper Kursat Uysal and Serkan Gunal 2014. The impact of preprocessing on text classification. Information Processing & Management 50,1(2014),104–112.
11. Jonathan J Webster and Chunyu Kit.(1992). Tokenization as the initial phase in NLP. In Proceedings of the 14th conference on Computational linguistics-Volume 4.
12. Hassan Saif, Miriam Fernández, Yulan He,and Harith Alani.(2014).On stop words, filtering and data sparsity for sentiment analysis of twitter (2014).
13. Julie B Lovins. 1968. Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory.

How to cite this article:

Sukhmani and Chauhan R.K.2018, A Novel Approach For Word Net Ontology Based Document Categorization Technique. *Int J Recent Sci Res.* 9(5), pp. 27191-27194. DOI: http://dx.doi.org/10.24327/ijrsr.2018.0905.2207

****