ISSN 2278-3083

International Journal of Science and Applied Information Technology (IJSAIT), Vol. 3 , No.4, Pages : 12 - 17 (2014)
Special Issue of ICECT 2014 - Held during September 01, 2014 in The Golkonda Hotel, Hyderabad, India

CSS-HDC: Hierarchical Document Classification by conceptual and semantic similarities

Venkata Ramana. A[1]Dr.E.Kesavulu Reddy[2]
[1]Research Scholar,
[1]Department of Computer Science
[1]S.V.University,Tirupat-India-517502
[1]Mobile: + 91 9441871400
[1]Email: avr_rdg@yahoo.co.in

[2]AMTIE(India).,M.C.A.,M.Phil(CS),PhD(CS)
[2]Assistant Professor
[2]Department of Computer Science
[2]S.V.University,Tirupat-India-517502
[2]Mobile: + 91 9866430097
[2]E-mail :ekreddysvu2008@gmail.com

Abstract: The victimization of syntactic components and semantic environment has constantly become a noteworthy issue in the milieu of data mining and information retrieval, which is in particular of text data. The effectiveness of this issue has delivered noticeably in absolutely unique tasks, such that as supervised learning of the text data. So significantly, still, extra syntactic or semantic info has become utilized only distinctively. With motivation gained from our earlier work that successfully able to define the concept labels for supervised learning, here in this paper we devise a hierarchical document categorization by conceptual and semantic relevance. The conceptual relevance is verified by concept labeling approach that devised in our earlier research article. Semantic relevance is explored by estimating the correlation between concept categories based on the activity labeling, which is main contribution of this paper. The results explored in empirical study concluding that the devised model is promising the significant classification by conceptual semantic relevance of given documents.

INTRODUCTION

Natural Language Processing (NLP) is actually a contemporary computational innovation also a technique of examining as well as determining phrases regarding human language itself. NLP is a term which connects back towards the traditions of Artificial Intelligence (AI), the basic review of intellectual function with computational activities, having a focus on the role of information illustrations. The demand for illustrations of human learning worldwide is needed to perceive human communication to computers.

Text mining efforts to explore newer, earlier not known data by employing methods from normal language process as well as data mining. Categorization, among conventional text mining strategies, is monitored learning perspective in which categorization techniques attempt to designate a doc to several classifications, according to the doc information. Classifiers are proficient from illustrations to perform the classification work instantly. To enhance efficient and effective understanding, every classification is addressed as a binary category challenge. The concern is whether a doc needs to be designated to a specific niche or not.

A lot of recent report categorization techniques are according to the vector space model (VSM) [1], [2] and [3] that is a commonly employed data description. The VSM signifies every doc as a characteristic vector of the jargon (words or phrases) in the doc. Every characteristic vector consists of term loads (usually term-frequencies) of the words in the doc. The resemblance in between documents is assessed by one of the resemblance actions that are according to that a characteristic vector. Illustrations consist of the cosine estimate as well as the Jaccard measure.

Generally, in text categorization strategies, the consistency of a term (word of phrase) is calculated to examine the significance of the term inside doc. Anyhow, two terms provide the equivalent consistency in a doc, yet one term brings much to the implying of its content as compared to the another term. So, a few terms supply the vital concepts in a conviction, and reveal such a conviction is all about. It's significant to notice that removing the interaction in between verbs as well as their arguments in the equivalent conviction has the prospective for evaluating terms inside a conviction. The details about who's performing what to whom explains the participation of every term in a conviction to the significance of the principal theme of that conviction.

The similarity determines shows the point of distance or splitting of the desired objects and might represent to the aspects that are suspected to identify the clusters enclosed in the data. Earlier Clustering, a similarity/distance assess should always be confirmed. [4]. Selecting an applicable similarity step is also significant for cluster assessment, particularly for a specific kind of clustering algorithms.

Text Categorization (TC) is the categorization of information with affection to a collection of one or additional preexisting

aspects [5]. The categorization phase includes of building a weighted vector for every aspect, and then applying a resemblance assesses to find the nearest category. The resemblance determines is used to identify the level of similarity in between two vectors. To accomplish reasonable categorization results, a similarity evaluate must usually respond with significant values to information that should be to the similar class and with modest values commonly. All Through the past decades, a huge number of techniques endorsed for text categorization had been commonly built on the traditional Bag-of-Words version where each and every term or term stem is a self-governing feature.

The prevailing similarity strategy was much more usually used to analyze the resemblance in between words. However the content theoretic likeness determines results are mathematically noticeable it does not diminish the specifications of the vector model [6].Metric ranges such that Euclidean distance is not really suitable for high specifications and sparse fields. Owning to the situation of any regards in between words, the learning algorithms are forbidden to identify patterns in the included terminology only, although conceptual patterns persist dismissed.

Prevailing strategies requires doing an optimization more than a whole assortment of documents. Many of these strategies are computationally expensive.

ASSOCIATED WORK

The improvements in this field are accelerated by tough theoretical motivations. This is because of the machine learning methods in the text classification field. For this classification a good number of machine learning techniques are used, which include example-based classifiers, neural networks, Rocchio method, nearest neighbor classifiers, regression methods, decision trees and probabilistic classifiers [7].

Vapnik introduced new learning methods Support Vector Machines (SVMs) in 1995 [8], [9]. Promising results were got in later years when many studied made use of SVMs for text classification [8-12], [13]. Joachims presented the primary studies that brought in SVMs for text classifications in 1998. the study shows the comparison of non-linear model with four popular machine learning algorithms Naïve Bayes (NB) classifier, Rocchio method, k-nearest neighbor (k-NN) classifier and C4.5 decision tree) on Reuters and Ohsumed datasets. We can conclude that SVM is perfect for test classification and most importantly is better than the other methods. Dumais et al in the same year checked the precision of five various machine learning algorithms on Reuters dataset for text classification and wrapped up with the result that the precision of the simple linear SVM is one of the best reported for Reuters alike the Joachims study. Linear SVM is specifically promising as it is much easy and more competent that Joachims non-linear model [10]. Yang and Liu conducted a controlled study with statistical important tests on five learning algorithms (SVM, k-NN, neural network (NNet) method, NB and Linear Least-Square Fit (LLSF) mapping). The conclusion was that SVM is one of the most successful machines to learn algorithms. A survey was present by Sebastiani which covered the main machine learning approaches in test classifications [11].

Text classification has another major issue which is decreasing dimensionality. By using feature selection we can obtain precision and efficiency of classifiers by choosing more discriminative terms in datasets as features. Different feature selection methods have been shown and checked in literature [14]. Five various feature selection methods were analyzed by Yang and Peterson, 1997 on Reuters and Ohsumed datasets by making use of k-NN and LLSF categorization algorithms in the case of global policy. IG and CHI methods are termed to be the most successful methods [15]. An empirical comparison was drawn by Forman. He compared twelve feature selection methods. This was done on a benchmark that was got from Reuters, TREC and Ohsumed, by making use of SVM in case of local policy. Outstanding performance was shown by accuracy and F-measure. Specifically on highly skew datasets yet it was IG that yielded the best results in exactness. Debole and Sebastiani in the same year suggested supervised term weighting (STW) scheme by making use of IG, CHI and gain ratios (GR) along with TF-IDF weighting on Reuters dataset along with SVM in both local and global policies. It was finalized that the GR performs better than the other methods and gave exemplary results as a STW function specifically in macro-averaged F-measure [16].

On Reuters data along with SVM in local and global policies Ozgur et al [17] compared tf-idf weighting with Boolean weighting. Compared to Boolean weighting it was observed that tf-idf performed better. It was also seen that global policy performed better for a large number of keywords than small number of keywords. In such case local policy outperformed the global policy [17]. In a study thereafter Ozgur and Gungor checked the performance of these two policies along with two weightings and six other document collections. Added to this there were skewed properties along with various numbers of keywords using SVM in detail in 2006. Also it was proven that the results of the earlier studies can simplify that the global policy is a better performer when it comes to large number of keywords and local policy performs well for small

number of keywords and in skew datasets [14]. Tasci and Gungor, 2006, used the analysis with different existing feature selection methods and four other suggested feature selection methods that are similar to Acc2 metric. The feature selection methods on six standard document collections were compared by changeable number of selected feature from 10 to 2000 in both local and global policy. Also they came to a conclusion that Acc2 is the finest metric among the existing metric, that too with a limited number of features. The victory of Acc2 was clear in local policy on skew datasets.

Contrary to this the suggested metric M1 is more winning that the victorious metric Acc2 in the experiments [18], [19].

Liu et al [20], centered on data imbalance issue in text classification by showing a probability based term weighing scheme that was stimulated by various feature selection approach. They wrapped up that making use of probability based term weighing scheme can perk up categorization performance on rare classes.

Different researches have been performed to better the working of feature selection approaches on text categorization. But these basically deal with the improvement of the performance of every single feature selection approach. Also, it is difficult to say which feature selection method is better than the other, even though there are many feature selection methods in text classification. In addition, the text is categorized based on the heaviness of the feature in relation to conceptualization but not in term relation. For example both coal mining and data mining can be put into one class called as mining, which is not accurate in the context of concept relation. Therefore in our earlier research article [21] defined a measuring metric called feature relationship weights that help us describe the class labels not by terms but by concepts.

CLASSIFICATION BY CONCEPTUAL AND SEMANTIC SIMILARITIES

The projected approach is a hierarchical supervised learning for text document categorization. The said model hierarchy is having two levels. The first level of the supervised learning is to classify the documents by concept and further in second level, these documents will be categorized by their semantic similarity. The semantic similarity of the documents will be identified by the correlation of the activities and concepts. In the first level of the said model is categorizing the documents by concept weights, which is measured from the feature correlation [21]. The second level of the projected model classifies these documents by the correlation of the activities found the given document text descriptions. Based on the NLP strategies the verbs used to bind the arguments are being considered as activities and the arguments are being considered as concepts.

Data Preprocessing

The initial step of the projected model is to preprocess the text data of given documents dataset. At first the text data will be tokenized and then the stop words will be removed. Further the leftover words will be processed by stemming, which mainly to remove tenses.Since these processed word tokens are used as input, the natural language process techniques are applied to identify the arguments as concepts and verbs as activities.

First level Categorization by Concept Labeling [21]

In the hierarchy of devised two levels supervised learning, categorizing the documents by concept labeling is the first level, which is based on our earlier work [21]. In regard to this, initially finds the concept weights and then the features are pruned through the metric called similarity score. The two approaches called measuring concept weights and pruning features by similarity score metric are briefed in following sections.

Concept Weights by Feature Correlation

The increasing order series of the features and their happening in the said set of documents is known as concept weight (cw). The ordered series is originally considered with lone feature and then it rises by adding every feature per iteration. The sequence is terminated once the concept weight is found to be less than the said threshold. In this process if the series s1 is subset of sequence s2 and concept weight of s1 is less than or equal to concept weight of s2 then s1 can be trimmed [46]. This procedure results set of hypothesis as feature set ‘CFS’

Feature pruning by similarity score

Feature with many terms will be got in this phase and the obtain the parallel score in each selected feature x and other each feature x’ with lesser number of terms than x. in case the semblance between x and x1 are found to be more than the said threshold and bigger than all of the similarity scores

ISSN 2278-3083

International Journal of Science and Applied Information Technology (IJSAIT), Vol. 3 , No.4, Pages : 12 - 17 (2014)
Special Issue of ICECT 2014 - Held during September 01, 2014 in The Golkonda Hotel, Hyderabad, India

between x' and rest of bigger length features selected then x1 will be grouped to the x.

The algorithmic approach of the feature set optimization in pseudo code format

Let ts be the terms set selected from CFS.

Let $c\omega_\tau$ be the concept weight threshold

Order the terms belongs to 'ts' in descending by their frequency score.

1. For each term $\{t_i \mid t_i \in ts\}$:

Begin

a. Let $\{c_i \mid c_i \in cs\}$

b. $c_i \leftarrow t_i$ (add t_i to c_i)

c. For each term $\{t_i^{'} \mid t_i^{'} \in ts, t_i^{'} \neq t_i\}$

Begin

Project concept $c_i^{'}$ by adding $t_i^{'}$ to c_i in sequence

If ($cw(c_i^{'}) \geq c\omega_\tau$)

If ($cw(c_i^{'}) \cong cw(c_i)$) then

Discard c_i;

Set $c_i \leftarrow c_i^{'}$ continue step c.

Else

$cs \leftarrow c_i^{'}$ (add $c_i^{'}$ to cs)

Continue step c;

Else

Discard $c_i^{'}$

Continue step c;

End of step c;

End of step 1.

Order cs in descending order by the length of the concepts (here after concepts referred as features)

Let ml be the maximal length of the feature in cs

For each $\{c_i \mid c_i \in cs; tc(c_i) \equiv ml\}$ move c_i to label set ls

2. For each $\{l_i \mid l_i \in ls; tc(l_i) \cong ml\}$;Here $tc(c_i)$ indicates the term count of the feature c_i

Begin

a. For each $\{c_i \mid c_i \in cs; c_i \notin ls\}$

Begin

Find similarity score

$$ss_{(l_i \leftrightarrow c_i)} = \frac{tc(l_i \cap c_i)}{tc(l_i \cup c_i)}$$

End of Step a;

End of Step 2;

3. For each $\{c_i \mid c_i \in cs; c_i \notin ls\}$

Begin

a. For each $\{l_j \mid l_j \in ls\}$ set $\bigcup_{j=1}^{|ls|} ss_{(l_j \leftrightarrow c_i)}$ in descending order and select first element as $ss_{l \leftrightarrow c_i}$

b. If ($ss_{l \leftrightarrow c_i} \geq ss_\tau$) then consider c_i as feature of the group represented by the label l

Else move c_i to ls

End of step 3;

If ls got updated then go to step 2 else

Return ls as set of class labels

Further the classification of the documents is initiated that performs supervised learning by using concept labels as the labels of the categories

Finding Correlation of the semantics

This stage of supervised learning estimates the correlation between activities that extracted from the given documents dataset. In this regard the activities found are considered to be categorical as they associate with divergent arguments.Henceforth here we use mean-square contingency coefficient [22] to estimate the correlation between attributes. Any given two activities A and B such that $\{a_1, a_2, a_3, \ldots..a_m\}$, $\{b_1, b_2, b_3, \ldots.b_n\}$ are categorical argumentsfound to be associated to A and B respectively. The size of the set of arguments associatedwithactivity A is m and activity B is n. Then the mean square contingency coefficient between activities A and B can be measured as follows:

$$\rho_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n} 1 - \frac{1}{o(a_i, b_j)}$$

Here in this equation ρ_{ij} is the fraction of co occurrence of a_i, b_j

$$\rho_i = \sum_{i=1}^{m} 1 - \frac{1}{o(a_i)}$$

Here in this equation ρ_i is the fraction of occurrence of a_i

$$\rho_j = \sum_{j=1}^{n} 1 - \frac{1}{o(b_j)}$$

Here in this equation ρ_j is the fraction of occurrence of b_j

International Journal of Science and Applied Information Technology (IJSAIT), Vol. 3 , No.4, Pages : 12 - 17 (2014)
Special Issue of ICECT 2014 - Held during September 01, 2014 in The Golkonda Hotel, Hyderabad, India

$$\chi^2_{(A\leftrightarrow B)} = \frac{1}{\min(m,n)-1} * \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(\rho_{ij} - (\rho_i.\rho_j)\right)^2}{\rho_i.\rho_j}$$

Here in this equation $\chi^2_{(A\leftrightarrow B)}$ is the mean square contingency coefficient that indicates the correlation between activities A and B.

According to the correlation estimation process explored here, the activities that are highly correlated will be grouped. Further each group of activities will be used as class label for second level of supervised learning.

Empirical Study

Dataset characteristics

The features of the datasets used in experiments have an important role. Two known text datasets namely Reuters-21578 [23] and 20 Newsgroups [24] are used in this experimental study, the details explored in Table 1. In these experiments Reuters -21578 ModApte split and 10 classes of 20 Newsgroup dataset are taken into consideration. The dataset of Reuters is skewed. Ever dataset will have a different number of documents. In contrast the dataset in the newsgroups will have even distribution with same number of documents in every class. Therefore the efficiency of these features can be seen in two different datasets with diverse features.

Performance Analysis

Table 1: Statistics of the experiment results

Total Number of documents	4136
Average of concepts in a document	132
Average number of Activities in a document	74
Total number of correlation concept label sets found	63
Total number of correlation activities found	31

Total number of documentsconsidered4136

Total number of documents found to be classified as false negative 32 and true negative 110)

Total number of documents found to be classified as true positives 3783 and false positives 211

As per the results explored above, the devised hierarchical supervised learning of documents is accurate to the level of 91.46%. The failure percentage is 8.53%, which is nominal.

The experiments also conducted on the same data set with earlier methodcalled optimizing features by correlating [21], which is not considering the semantic similarities of the features, and the results are as follows:

Total records Tested 4136

Total number of documents found to be classified as false negative 550 and true negatives 670

Total number of documents found to be classified as false positives 534 and true positives 2382

As per these results, the accuracy of the earlier devised model [21] is less significant since we observed that the prediction success limited to 57.59%. The failure percentage is approx 42.5%, which is not a negligible factor.

Hence it is obvious to conclude that the semantic similarity along with concept similarity scoreis more significant compared to alone concept similarity towards the supervised learning (see fig 1).

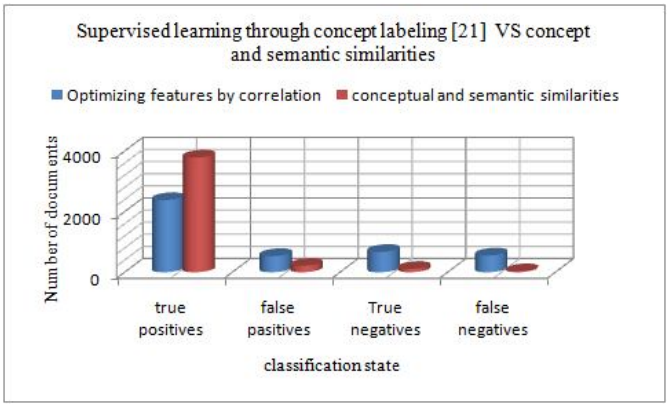

Fig 1: Performance analysis of supervised learning through concept labeling versus concept and semantic similarities.

CONCLUSION

Here in this paper, a novel hierarchical supervised learning strategy has been devised. The said model is with two levels of document classification, which in first level classifies the documents by estimating the feature correlation through concept labeling and then in second level classifies these classified groups again by estimating the semantic similarities. The experimental results explored here are indicating the significance of the devised hierarchical supervised learning is miles ahead of the supervised learning model devised in our earlier work [21]. In future research the said model can be implemented even in unsupervised learning strategies.

REFERENCES

[1] K. Aas and L. Eikvil. Text categorisation: A survey. technical report 941. Technical report, Norwegian Computing Center, June 1999.

[2] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):112–117, 1975.

[3] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[4] Similarity Measures For Text Document Clustering Anna Huang 2008.

[5] F. Sebastiani. Machine Learning In Automated Text Categorization. Acm Computing Surveys, 34(1):1–47, 2002.

[6] S. ClinchantAnd E. Gaussier. Information-Based Models For Ad Hoc Ir. Proceedings Of 33rd Annual International AcmSigir Conference On Research And Development In Information Retrieval, Pages 234–241, 2010.

[7] FabrizoSebastiani, machine learning in automated text categorization ACM computing surveys, Vol.34,No 1,March 2002, pp.1-47.,2002.

[8] Tang, N., “Text Categorization using Support Vector Machines”, 2001

[9] Vapnik, V., “Statistical Learning Theory”, AT&T Labs-Research, London University, Wiley, 1998.

[10] Dumais, S. T., J. Platt, D. Heckerman and M. Sahami, “Inductive learning algorithms and representations for text categorization”, Submitted for publication, 1998.

[11] Yang, Y. and X. Liu, “A Re-examination of Text Categorization Methods”, Proceedings of the Twenty-Second International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 42-49, 1999.

[12] Yizhang, G., “Methods for Patterm Classification”, New Advances in Machine Learning, InTech Press, pp. 49-74, 2010

[13] Cooley, R., “Classification of news stories using support vector machines”, IJCAI'99 Workshop on Text Mining, Stockholm, Sweden, 1999.

[14] Ozgur, A. and T. Gungor, “Classification of Skewed and Homogenous Document Corpora with Class-Based and Corpus-Based Keywords”, 29th German Conference on Artificial Intelligence (KI 2006), Bremen - LNAI (Lecture Notes in Artificial Intelligence), Vol.4314, pp.91-101, Springer-Verlag, 2006.

[15] Y. Yang, J.O. Pedersen, “A comparative study on feature selection in text categorization”, Proceedings of the 14th International Conference on Machine Learning, pp. 412–420, 1997.

[16] Debole, F. And F. Sebastiani, “Supervised Term Weighting for Automated Text Categorization”, Proceedings of SAC-03-18th ACM Symposium on Applied Computing, ACM Press, pp. 784–788, 2003.

[17] Ozgur, A., L. Ozgur and T. Gungor, “Text Categorization with Class-Based and Corpus-Based Keyword Selection”, Proceedings of the 20th International Symposium on Computer and Information Sciences, Lecture Notes in Computer Science, Vol.3733, pp.607-616, Springer-Verlag, 2005.

[18] Tasci, S., “An evaluation of existing and new feature selection metrics in text categorization”, Computer Engineering, Bogaziçi University, 2006

[19] Tasci, S. and T. Gungor “An evaluation of existing and new feature selection metrics in text categorization”, Computer and Information Sciences ISCIS '08 23rd International Symposium on, pp.1-6, 2008

[20] Liu, Y., H. T. Loh and A. Sun, “Imbalanced text classification: A term weighting approach”, Expert Systems with Applications, Volume: 36, Issue: 1, Elsevier Ltd, pp.690-701, 2009

[21] Venkata Ramana, A; Naidu, M.M., "Optimizing features by correlating for concept labeling in text classification," Advance Computing Conference (IACC), 2014 IEEE International ,vol., no., pp.561,567, 21-22 Feb. 2014; doi: 10.1109/IAdCC.2014.6779386

[22] G. Parthiban, A. Rajesh, S.K.Srivatsa “Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method”

[23] https://archive.ics.uci.edu/ml/machine-learning-databases/reuters21578-mld/

[24] http://kdd.ics.uci.edu/databases/20newsgroups/20_newsgroups.tar.gz