Semantic Document Clustering using Recurrent Neural Network

Priyanka B. Sonawane[1], Prof. Pramila M Chawan[2]

[1]M.Tech Student, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India
[2] Associate Professor, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

---

Abstract: *Class-label classification is an important machine learning task where in one assigns a subset of candidate without label to an object. In this paper, we propose a new document classification method based on Machine Learning (ML) approach. Our proposed method has several attractive properties: it captures some metadata from each object and built the train the training set first. Each object has categorized in under the specific class label according to achieved weight. Recurrent Neural network (RNN) has used for proposed classification model. The IEEE documents has used for training and testing purpose, which is distributed according to cross fold validation. The 70-30% data has taken for training and testing respectively. Some partial implementation demonstrates the results of system in results chapter. The experimental analysis shows the effectiveness of proposed system than traditional document classification approaches.*

Keywords: Document classification, RNN, Deep Learning, Multi label classification.

Introduction

Classification is very essential in data mining as well as machine learning approaches. Now days many sources has generates the different types of data in row format, and its hard to process from current environments and algorithms also. Text classification is to map the text to one or more predefined categories using a kind of classification algorithm which is accomplished according to text content. A standard classification corpus has been established and a unified evaluation method is adopted to classify English text based on machine learning which has made a large progress now. Most real world data are stored in relational databases. Document clustering is an important machine learning task wherein one assigns a subset of candidate labels to an object, the main issue of multi label clustering is the redundant clustering approach for online as well as offline data set to handle this issue, we have planned to use density based re-clustering of existing micro-clustering objects and improve the maximize accuracy of final sub-clusters. Demonstrate two implementations of our method using logistic regressions and gradient boosted trees, together with a simple training procedure based on Expectation Maximization. We further derive an efficient prediction procedure based on dynamic programming, thus avoiding the cost of examining an exponential number of potential label subsets. For the testing, we will use and show the effectiveness of the proposed method against competitive alternatives on benchmark data sets with pdf. An increasing number of data mining tasks includes the analysis of complex and structured types of data and make use of expressive pattern languages. Most of these applications can't be solved using traditional data mining algorithms. This work address the issue redundant document clustering and eliminate it using proposed algorithms. The different pdf dataset has used for testing and create the runtime micro cluster as well as IEEE dataset has used for generating the Background Knowledge (BK) of system.

Literature Survey

According to Lubomir Stanchev et. Al. [1] has represented Semantic Document Clustering Using Information from WordNet and DBPedia. this technique can group documents that share

no words in common as long as they are on the same subject. We compute the similarity between two documents as a function of the semantic similarity between the words and phrases in the documents. We model information from WordNet and DBPedia as a probabilistic graph that can be used to compute the similarity between two terms. The Cosine Similarity (CS) algorithm has used for generate the similarity weight between two vectors, once apply ordering algorithm on sorted vectors system will generate the classified results.

Gupta, Aditi, Jyoti Gautam et. Al. [2] proposed a A Survey on Methodologies used for Semantic Document Clustering. This approach describes the survey of various research papers that have been studied and highlights the merits and demerits of each clustering algorithm. This will give a direction to future research in a more focused manner.

Kulathunga, Chalitha et. Al [3] proposed a An ontology based and domain specific clustering methodology for financial documents. The Relational Data Framework (RDF) framework has used to extract the semantic knowledge. It is generally used for identifying the correct meanings of the ambiguous words in the documents. Most of the proposed methodologies were experimented on general document datasets and most of the few available domain specific clustering studies were constrained to specific domains where complete domain ontologies are available. Although financial domain has several domain ontologies, none of them are complete and suitable for semantic document clustering

Radhika K, Bindu K.R.et. Al. [4] has presented A Text Classification Model Using Convolution Neural Network and Recurrent Neural Network. It proposed text classification model using CNN and RNN. Text classification is defined as categorizing document into one of the category in which the text belongs to Neural Networks are used for classification. Collection of documents is trained and tested using neural networks. In this paper we build a text classification model using Convolution Neural Network and Recurrent Neural Network. We train and test both CNN and RNN model with our dataset. The dataset we used is collection of essays. From the train and test accuracy obtained we reach a conclusion that RNN performs better than CNN for our essay dataset.

Pengfei Liu, Xipeng Qiu, Xuanjing Huang et. Al. [5] has Presented Recurrent Neural Network for Text Classification with Multi-Task Learning. Here we use the multitask learning framework to jointly learn across multiple related tasks. Based on recurrent neural network, we propose three different mechanism of sharing information to model text with task-specific and shared layers. The entire network is trained jointly on all these tasks. Experiments on four benchmark text classification tasks show that our proposed models can improve the performance of a task with the help of other related tasks. It also describe the multitask learning framework to jointly learn across multiple related task based on RNN.

Guimaraes, Rita Georgina, et al. [6] presented in the Age Groups Classification in Social Network Using Deep Learning. This research suggests that one of the most relevant parameter contained in the user profile is the age group, showing that there are typical behaviors among users of the same age group, specifically, when these users write about the same topic. A detailed analysis with 7000 sentences was performed to determine which characteristics are relevant, such as, the use of punctuation, number of characters, media sharing, topics, among others; and which ones can be disregarded for the age groups classification. Different learning machine algorithms are tested for the classification of the teenager and adult age group, and the Deep

Convolution Neural Network (DCNN) had the best performance, reaching a precision of 0.95 in the validation tests

Proposed System

In the project we are using standard IEEE as well as some document for training as well testing purpose. For the training and testing criteria has been given 70/30 (Training/Testing). Below modules detail shows linear execution of system.

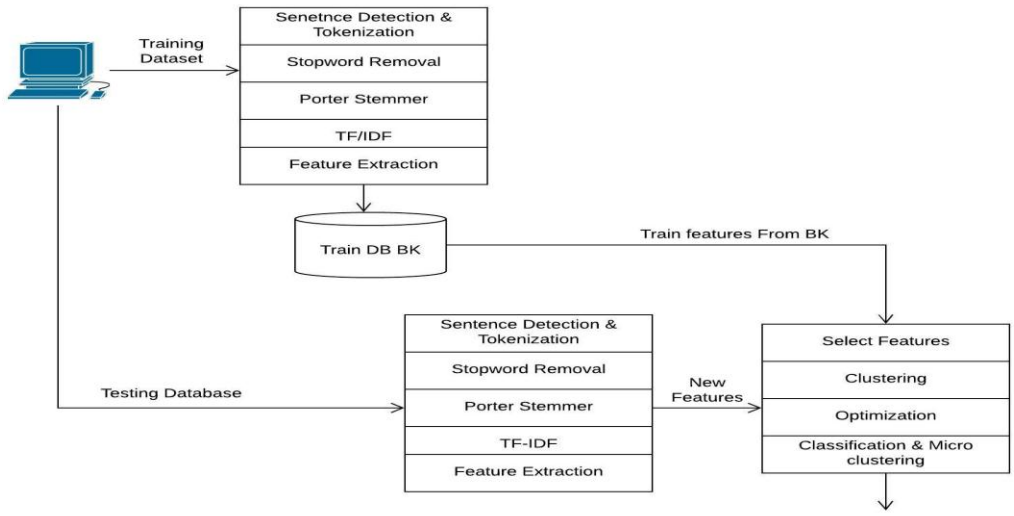

Figure 1 : Proposed system architecture

The above Figure 1 illustrates the proposed system execution with training as well as testing module. It describes the machine learning and some data mining strategies, below are the system phases which are carried out for overall proceeding.

Data Training phase with pre-processing

This module performs data pre processing to create train dataset.

Then first upload the training directory of pdf dataset.

Once upload it will read the data from PDF using PDFBOX API.

Then tokenization, stop word removal and porter's stemmer will execute.

Finally TF-IDF will provide the availability of current vector and store into feature database

Testing phase with preprocessing and TF-IDF

First upload the test directory of pdf as well image dataset.

The initial phase of testing is same like training phase till IDF score calculation.

Then features are extracted using RNN

And classification is done using similarity vector

Feature Selection phase

This module extracts the feature form all buckets using Optimization approach.

Initial pheromone need to set.

The pheromone will select the neighbours and strong node for selection.

RNN Classification module

Here RNN use for classification purpose.

Here we find the training dataset with domain detail and feature details.

Once RNN execute it will ask for variation as well generation, after that optimized the results.

Finally similarity score will classify each bucket into the respective domain.

Algorithms

1 : Stop word Removal Approach

Input: Stop words list L[], String Data D for remove the stop words.

Output: Verified data D with removal all stop words.

Step 1: Initialize the data string S[].

Step 2: initialize a=0,k=0

Step 3: for each(read a to L)

If(a.equals(L[i]))

Then Remove S[k]

End for

Step 4: add S to D.

Step 5: End Procedure

2 Stemming Algorithm.
Input : Word w

Output : w with removing past participles as well.

Step 1: Initialize w

Step 2: Intialize all steps of Porter stemmer

Step 3: for each (Char ch from w)

If(ch.count==w.length()) && (ch.equals(e))

Remove ch from(w)

Step 4: if(ch.endswith(ed))

Remove 'ed' from(w)

Step 5: k=w.length()

If(k (char) to k-3 .equals(tion))

Replace w with te.

Step 6: end procedure

3 TF-IDF
Input : Each word from vector as Term T, All vectors V[i...n]
Output : TF-IDF weight for each T
Step 1 : Vector = {c1, c2, c3….cn}
Step 2 : Aspects available in each comment
Step 3 : D = {cmt1, cmt2, cmt3, cmtn}
and comments available in each document
Calculate the Tf score as
Step 4 : tf (t,d) = (t,d)
t=specific term
d= specific document in a term is to be found.
Step 5 : idf = t → sum(d)
Step 6: Return tf *idf

Results and Discussion

The implementation of proposed system has been completed for the training module. As per our first module we have used standard pdf data set of 100 files for training. The below table shows the training data.

Table I: Training dataset

Sr. no	No. of Papers	Label	Domain Name
1	15	1	Data Mining
2	15	2	Machine Learning
3	15	3	Cloud Computing
4	15	4	Network Security
5	15	5	Soft computing
6	15	6	Artificial Intelligence
7	10	7	Image Processing

After uploading dataset system reads abstract section using PDFBOX API. Then we preprocessed the data by using algorithms stop word removal and porter stemmer. Then TF-IDF stores the features vector. For this system, the system performance evaluation, will calculate the confusion matrix for accuracy. Then estimate classification results for different test data size.

The below analysis is the system classification graph. The graphs display how system classify the overall inputs into categories. The proposed system is implemented with RNN combination, which gives all results with satisfactory level. For performance evaluation 112 documents given for training and 30 documents given for testing. Here system compares the proposed results with two different existing systems.

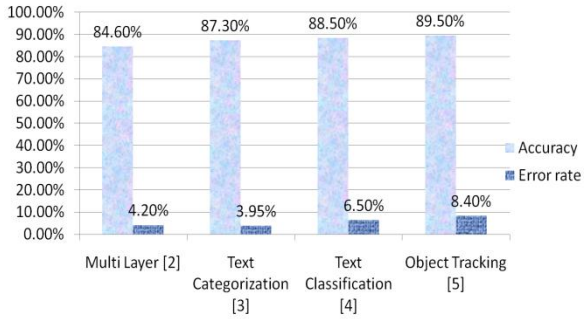

Figure 2: Domain Classification Accuracy

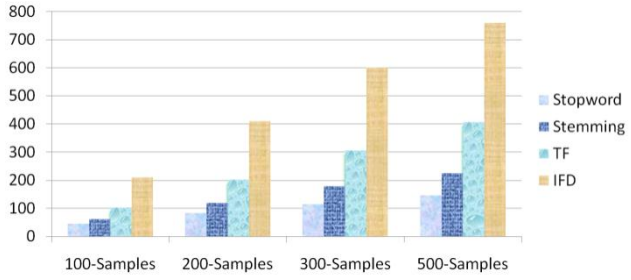

Figure 3 : NLP process execution time with number of samples

The above figure 3 shows time required for each processing according to different samples, the time has given based on seconds for desired system configuration.

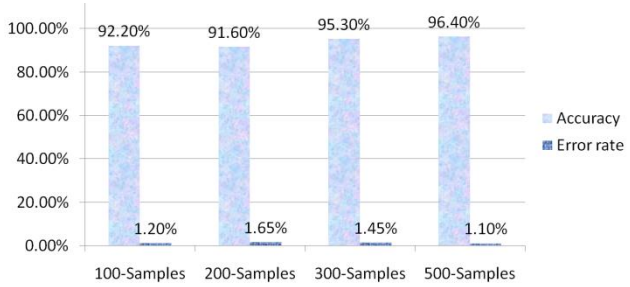

Figure 4 : Accuracy of proposed system with various samples

Conclusion

In this work proposed a document clustering and multilabel classification using RNN approach. In this work proposed a document clustering and multilabel classification using RNN approach. The proposed learning scheme explicitly models the multi-label correlations by label graph learning, which is jointly optimized with multilabel classification. As a result, the learned label correlation graph is capable of well-fitting the multi-label classification task while effectively reflecting the underlying topological structures among labels. In addition, we have presented a community-aware regularize to capture the context-dependent inter-label interaction information. The proposed work can classify the strong label with test instance using NN weight calculation as well classification approach.. Experimental results have demonstrated the effectiveness of our approach over several benchmark datasets.

Future Work

Sometime system having a accuracy issues well false detection ratio, we can focus on such problems. The second part is system execution complexity when we work with high dimensional or big data. The system can be work with HDFS framework for minimum time

computation or parallel distribution. So For the enhancement system can be execute HDFS base architecture with parallel genetic algorithm.

References

[1] Gupta, Aditi, Jyoti Gautam, and Ajay Kumar. "A survey on methodologies used for semantic document clustering." 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). IEEE, 2017.

[2] Kulathunga, Chalitha, and D. D. Karunaratne. "An ontologybased and domain specific clustering methodology for financial documents." Advances in ICT for Emerging Regions (ICTer), 2017 Seventeenth International Conference on. IEEE, 2017.

[3] Nema, Puneet, and Vivek Sharma. "Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique." Computers, Communications, and Systems (ICCCS), International Conference on. IEEE, 2015.

[4] Pratama, Timothy, and Ayu Purwarianti. "Topic classification and clustering on Indonesian complaint tweets for bandung government using supervised and unsupervised learning." Advanced Informatics,
Concepts, Theory, and Applications (ICAICTA), 2017 International Conference on. IEEE, 2017.

[5] Guimaraes RG, Rosa RL, De Gaetano D, Rodriguez DZ, Bressan G. Age Groups Classification in Social Network Using Deep Learning. IEEE Access. 2017;5:10805-16.

[6] Kulathunga C, Karunaratne DD. An ontology-based and domain specific clustering methodology for financial documents. InAdvances in ICT for Emerging Regions (ICTer), 2017 Seventeenth International Conference on 2017 Sep 6 (pp. 1-8). IEEE.