ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 6, June 2016

A Survey on Different Techniques of Classification of News and Research Articles

Suraj Prasad, Prof.Manaswini Panigrahi

Abstract— **As the digital data increases on servers different researcher have focused on this field. As various issues are arise on the server such as data handling, security, maintenance, etc. In this paper text classification study is done which brief various techniques of classification with there implementations. Here different features for the text classification is explained in detailed with there requirements as feature vary as per text analysis. Paper has brief different evaluation parameters for the study and comparison of classification techniques.**

Index Terms— **Classification analysis,** Supervised Classification, Un-supervised Classification, Text Feature, Text Mining, Text Ontology,

I. Introduction

Text Mining [1] is the discovery by computer of new previously unknown information by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar within web search. In search user is looking for something already known and has been written by someone else.

Text Mining offers a solution to this problem by replacing or supplementing the human reader with automatic system Undeterred by the text explosion. It involves analyzing of documents to discover previously unknown information. The information might be relationship or pattern that are buried in the document collection and which would otherwise be extremely difficult, if not possible, to discover text mining can be used to analyzed natural language documents about any subject, although much of the interest at present is coming from biological science. Originally, research in text categorization analyzed binary problem, where a document is either relevant or not. Text mining involves the application of technique from are as information retrieval, natural language processing, data mining and information extraction.

Information retrieval (IR)) system identify the documents in a collection which match a user's query. The most well known IR system are search engine as google, Which identify those documents on the world wide web that are relevant to a set of given words IR systems are often used in libraries, where the documents are typically not the book themselves but digital record containing information about the books. this is however changing with the advent of digital libraries, where the document being retrieved are digital version of books and journals IR system allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally intensive algorithm to large document collection, IR can speed up the analysis considerably by reducing documents for analysis. for eg. if we interested in mining information only about protein interaction, we might restrict our analysis to documents that contains the name of a protein or some form of the web 'to interact' or one of its synonymous.

II. Features of Text Mining

1) Title feature

The word in sentence that also occurs in title gives high score. This is determined by counting the number of matches between the content word in a sentence and word in the title. In [4] calculate the score for this feature which is the ratio of number of words in the sentence that occur in the title over the number of words in the title.

2) Sentence Length

This features is useful to filter out short sentence such as datelines and author names commonly found in the news articles the short sentences are not expected to belongs to the summary. In [5] use the length of sentence, which is the ratio of the number of words occurring in the sentence over the words occurring in the longest sentence of the documents.

3) Term Weight

The frequency of the term occurrence with a documents has been used for calculating the importance of sentence. The score of a sentence can be calculated as the sum of the score of words the sentences. The score of important score wi of word i can be calculated by traditional tf.idf method.

4) Sentence position

Whether it is the first 5 sentence in the paragraph, sentence position in text gives the importance of the sentences. This features can involve several items such as the position of the sentence in the documents, section and the paragraph, etc, proposed the first sentence of highest ranking. The score for this features in [6] consider the first 5 sentence in the paragraph.

5) Sentence to sentence similarity

This feature is a similarity between sentences for each sentence S , the similarity between S and each other sentence is computed by the cosine similarity measure with a resulting value between 0 and 1 [6]. The term Weight wi and wj of term t to n term in sentences Si and Sj are represented as the vector. The similarity of each sentence pair is calculated based on similarity.

6) Proper Noun

The sentence that contains more proper noun (name entity is an important and is most probably include in the document summary . The score for this feature is calculate as the ratio of the number of proper noun that occur in the sentence, over the sentence length.

S_f(6)S = No. Proper noun in S/Sentence Length (S)

7) Thematic Word

The number of thematic word in the sentence, this feature is important because term that occurred frequently in a document

are probably related to the topic. The number of thematic word

indicates the word with maximum possible relativity. We used the top 10 most frequent content word for consideration as thematic. the score for this features is calculated as the ratio of the number of thematic words that occure in the sentence over the maximum summary of thematic word in the sentence.

S_f7(S) = No.thematic word in S/Max(No.thematic word)

III. Techniques of Classification

KNN (K Nearest Neighbors algorithm) in [4] is used which utilize nearest neighbor property among the items. This algorithm is easy to implement with high validity and required no prior training parameters. Although K nearest neighbor is also identified as instance based learning in other words classification of items is quite slow. In this classification techniques distance between the K cluster center and classifying item is calculated then assign item to cluster having minimum distance from the cluster center. In case of text mining features from the document is extracted then k labeled node is select randomly which are suppose to be cluster center and rest of nodes or document are unlabeled nodes. Finally distance between labeled and unlabeled node is calculate on the base of feature vector similarity. In this algorithm distance between nodes are estimate in log(k) time

Author	**Technique**	**Merit**	**De-merit**
Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu. 2012	Create Ontology by using term feature of the document.	Classify on the basis of keywords present in research papers.	Less efficient as pattern based classification work well.
Souneil Park, Jungil Kim, Kyung Soon Lee, and Junehwa Song. 2013	Disputant and document classification is done by modified version of HITS algorithm and an SVM classifier.	Efficiently classify documents on the basis of disputant sentences.	Need prior knowledge for disputant identification.
Esra Saraç, Selma Ayşe Özel. 2013	Utilize fire fly genetic approach to classify documents.	It require less execution time.	Classification accuracy is quite less.
Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana. 2015	This work adopt term as well [attern feature for classifying document in two category.	Utilization of both feature increase the classification accuracy.	Process required high execution time.

Advantages: Main significance of this algorithm is that this is robust against raw data which contain noise. In this algorithm prior training is not required as done in most of the neural network for classification. One more flexibility of this algorithm is that this work well in two or multiclass partition.

Advantages: Main significance of the Support Vector Machines is that it is less susceptible for over fitting of the feature input from the input items, this is because SVM is independent of feature space. Here classification accuracy

Limitations: In this work selection of appropriate neighbor is quite high if population of item is large in number. One more issue is that it required much time for finding the similarity between the document features. Because of these limitations this algorithm is not practical with large number of items. So cost of classification increases with increase in number of items.

Support Vector Machine (SVM) in [3] is quite famous soft computing technique for item classification which is based on the input feature vector quality and training of the support vector machine. In this technique an hyperplane is build between the items this hyperplane classify the items into binary or multi class. In order to find the hyperplane equation is written as P = B+XxW where X ia an item to be classify then W is vector while B is constant. Here W and B is obtained by the training of SVM. So SVM can perfectly classify binary items by using that calculated hyperplane.

Limitations: In this classification multiclass items are not perfectly classify as number of items reduce gap of hyperplane.

Image Classification

Fuzzy classification in [15], has classify image data which is highly complex and required stochastic relations for the creation of feature vector from images. Here different types of relations are combined where members of the feature vector is fuzzy in nature. So this relation based image classification is highly depend on the type of image format as well as on the threshold selection.

Advantages: This algorithm is easy to handle, while stochastic relation help in identifying the different uncertainity properties.

Limitation: Here deep study is required to develop those stochastic relation, accuracy is depend on prior knowledge.

IV. Preprocessing

As document is collection of paragraphs. Paragraphs are collection of sentences. While sentences are collection of words. So whole preprocessing focus on word in the document without any punctuations. So in pre-processing of document there are two common steps first is stop word removal, and second is stem word removal. [8]

Stop List Removals: As sentence is frame with number of words but some of those words are just use to contruct a proper sentence although it does not make any information in the sentence. So identification of those words then removing is term as Stop word removal. So a list of words is store by the researcher which help in identifying of stop words. This removal of stop words help in reduce the execution time of the algorithm, at the same time noisy words which not give any fruitful information is also removed. Stop words are like {a, the, for, an, of, and, etc.}. So text document is transform into collection of words which is then compare with these words and then each match word is removed from the document.

Inorder to understand this assume an sentence {India is a great country in the world} then after pre-processing it become {India, great, country, world} while stop words {is, a, in, the} in the sentence are removed. Let

Stem Word Removal In this words which are almost similar in prefix are replace by one word. This can be said collection of words share same word is term as stem. So there occurrence in the document make same effect but while processing in text mining algorithm it make different so update each word from the collection into single word is done in this stem word removal pre-processing step. Let us assume an collection of words for better understanding of this work. Collection of word is {play, plays, playing} then replace each with word {play}.

V. Evaluation Parameter

As various techniques evolve different steps of working for classifying document into appropriate category. So it is highly required that proposed techniques or existing work need to be compare on same dataset. But document cluster which are obtained as output is need to be evaluate on the function or formula. So following are some of the evaluation formula which help to judge the classification techniques ranking.

Precision=True positive/ (True positive + False positives)

Recall=True positives/ (True positive + False negative)

F-score= 2*Precision*Recall/ (Precision + Recall)

Accuracy = (True Positive + True Negative) / (True Positive + True Negative+ False Positive + False Negative)

In order to evaluate result there are many parameter such as accuracy, precision, recall, F-score, etc. obtaining values can be put into the mentioned formula to get better result.

.

.

CONCLUSION

As the writing work of different articles from laboratory, organization, press media, institutes are increasing day by day. Then publishing their work is also increase which is done by most of the journals , news paper, organizations. Here paper has cover an important issue of document classification. Various techniques with there required features are discussed in detailed. Here paper related work of researchers done in this field. So it can be concluded that one strong algorithm is required that can effectively classify document while it need an strong ontology for same. .

REFERENCES

1. Selma Ayşe Özel. Esra Saraç " Web Page Classification Using Firefly Optimization ", 978-1-4799-0661-1/13/$31.00 ©2013 Ieee.
2. S. Somasundaran And J. Wiebe, "Recognizing Stances In Ideological Online Debates," Proc. Naacl Hlt Workshop Computational Approaches Analysis And Generation Emotion In Text (Caaget '10), Pp. 116-124, 2010.
3. G. Salton, C. Buckley, "Term-Weighting Approaches In Automatic Text Retrieval" Information Processing And Management 24, 2008. 513-523.
4. L. Suanmali, N. Salim, M.S. Binwahlan, "Srl-Gsm: A Hybrid Approach Based On Semantic Role Labeling And General Statistic Method For Text Summarization", Research Article- Journal Of Applied Science, 2010.
5. M. K. Dalal, M. A. Zaveri, "Semisupervised Learning Based Opinion Summarization And Classification For Online Product Reviews",
6. Hindawi Publishing Corporation Applied Computational Intelligence And Soft Computing, Volume 2013.
7. A. Kiani, M. R. Akbarzadeh, "Automatic Text Summarization Using: Hybrid Fuzzy Ga-Gp", International Conference On Fuzzy Systems, 2006 Ieee. [11] Base Paper
8. Ning Zhong, Yuefeng Li, And Sheng-Tang Wu "Effective Pattern Discovery For Text Mining". Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012
9. Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, And Ou Liu. "An Ontology-Based Text-Mining Method To Cluster Proposals For Research Project Selection". Ieee Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012
10. Disputant Relation-Based Classification For Contrasting Opposing Views Of Contentious News Issues Souneil Park, Jungil Kim, Kyung Soon Lee, And Junehwa Song. Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013.
11. S. Park, K.S. Lee, And J. Song, "Contrasting Opposing Views Of Contentious Issues," Proc. 49th Ann. Meeting Assoc. Computational Linguistics (Acl '11), Pp. 340-349, 2011.

Author 1

Suraj Prasad is persuing M tech fro m IES's college of Technology Bhopal, he has completed BE from Millennium Inst.. of Technology Bhopal in 2014,he has attened many seminar in C/C++,java ,php,and awarded from so many institutions,

He has knowledge in data mining ,compiler design,theory of computation,operating system,networking, his teaching style is just like a introspection and heuristic ,

He is a man of letter and he believe in hard work .

M-8962181499

Author 2

Prof.Manaswini panigrahi is a versatile faculty of cse department in IES's college,she has completed M tech from North Orissa university in 2013 and her interested field is neural network,soft computing,data mining,pattern recognition,under her guidance a huge no.of paper has published,
She has 2 year teaching experience.
M=9644824532