A Novel Approach to Improve Efficient Information Retrieval based on Natural Language Processing

R.Umamaheswari
Assistant Professor/CSE, Gnanamani College of Technology,Namakkal, T.N, India

Dr.N.Shanthi
Professor & Dean, Nandha Engineering College, Erode, T.N, India

M.Thenmozhi
PG Scholar-M.E(CSE), Gnanamani College of Technology, Namakkal, T.N, India.

Abstract : Today, the web documents are available in electronic form and also it is scattered all over the world. The finding of exact information is very difficult. The Natural Language Processing (NLP) is used to build the good representations of meaning from unstructured text information. The classification process also plays important role while extract the information. In our proposed work the text processing and the Term Vector is constructed using Wordnet based relations. Wordnet provides synonyms, hyponyms and hypernym based relations for the given document set. This method gives 0.5% accuracy over other method. The emerging areas like data mining, web searching, Information retrieval both structured and unstructured data has lots of challenges. To evaluate the good pattern classification and clustering techniques are used. NLTK tool kit is used for text categorization. For good classification and semantic related classifications NLP (Natural Language Processing) concepts is used. Semantic similarity is the measure used to extract information based on concept wise and context wise. Most of the information scattered in the form of text in World Wide Web (WWW). NLTK provides good analysis in semantic similarity between sentences and words. The new architecture is proposed for improving the efficient searching and information retrieval.

Keywords: Information Retrieval, Classification, Natural Language Processing.(NLP).

I. INTRODUCTION

Text mining refers to the process of deriving high quality information from World Wide Web. Information are retrieved by using statistical learning methods. The data contained in Natural Language Text. The natural language text contains ambiguous data because of syntax and semantics. Text mining techniques converts the text document into numerical values for analysis. In this proposed work the various machine learning algorithms were studied related to Text Mining. Web mining is one of the applications of data mining. The Web Mining can be categorized into Web usage mining, Web content mining and web structure mining. Our work concentrated on web content mining. The web content mining mainly used for Information Retrieval. The documents are represented as Vector Space Model. To analyse the importance of the word Term Frequency and Inverse document Frequency (TF-IDF) is used. The usual evaluation metrics are precision and recall. In this paper we studied about the various text mining algorithms and different types of measures used for text classification.

Semantic similarity and semantic relatedness is used for extracting information based on meaning of text. Similarity example “Car is like bus”, whereas semantic relatedness includes relation between two words using word net. In this study work we analysed the learning algorithms and its pros and cons, finally we proposed the new architecture for text classification. Different tools and techniques were studied and analysed finally we identified the NLTK (Natural Language Processing Tool Kit) performs well for the text classification. In this tools there are many built in algorithms embedded and which helps for good classification. The Syntax and semantic oriented structure analysis can be done easily.

II. RELATED WORK

In [3] most of the paper given the statistics about text spread over the World Wide Web. The 80% of the content is in unstructured text. They concentrated proper annotation of text document, presentation of text

document, and classification.Natural Language Processing (NLP) is to achieve a better understanding of natural language by use of computers and represent the documents semantically to improve the classification and informational retrieval process.Pratiksha Y. Pawar,Gawandeintroduced various text categorization methods, for getting efficient Information retrieval systems. They used supervised learning techniques. The categorization is based on predefined categories. Text classification is done by using machine learning algorithms. The main drawback of this method is supervised method consumes times during model construction. Here it is proved that hybrid approaches gives better results than the individual machine learning algorithms. The SVM (Support Vector Machines) classifier is recognized as most efficient classifier.In [7] this paper proposes text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the different types of the documents like web kb, newsgroups, Reuters etc., They mainly concentrated single label class.In [9] they proposed the combined method for calculating semantic similarity between words are used. This paper combines page count based and lexicon-syntactic patterns extracted from snippets to leverage a robust web based semantic similarity measure. The SVM classification is used for document classification. They proposed lexicon-syntactic patterns based approach to compute semantic similarity using snippets obtained from a web search engine. They integrated different web based similarity measures using WordNetsynsets and support vector machines to create a robust semantic similarity measure. The Charles – Miller bench mark datasets were used. In that work they combined both WordNetsynsets and web content to leverage a robust semantic similarity measure. Main drawbacks of to find synonyms from the web.

In Sentence Similarity [SS] computes a similarity score between two sentences [10]. The SS task differs from document level semantics tasks in that it features the sparsely of words in a data unit, in this paper, they hypothesized that by better modeling lexical semantics. They incorporated both corpus-based (selection preference information) and knowledge-based (similar words extracted in a dictionary) lexical semantics into a latent variable model.
In[2] clustering is a useful technique that organizes a large quantity of unstructured text into a smaller number of groups, which helps for information retrieval. K-means algorithms works better for large datasets than hierarchical clustering algorithms. They used various distance measure for calculating similarity measures like Euclidian distance, cosine distance, Jacquard co-efficient and Pearson co-efficient. They analysed different types of datasets and its entropy results. In [1] they focused the density of clusters. The analysis has given for comparing the cluster distance. Based on their analysis, they concluded that Euclidean function works best with SNN clustering approach in contrast to cosine, Jacquard and correlation distance measures function. This method works well for more number of outliers and noises. Future work analyse impacts of other different similarity measure functions upon various popular clustering techniques.
In [8] Semantic similarity can be broadly construed between texts of any size. Depending on the granularity of the texts, we can talk about the following fundamental text-to-text similarity problems: word-to-word similarity, phrase-to phrase similarity, sentence-to-sentence similarity, paragraph-to-paragraph similarity, or document- to-document similarity.
All existing system has poor document representation, so the semantic similarity is poor. It handles large amount of data, and this leads to poor understanding the system. There are several text mining or extract information are available, classification accuracy is less.

III. PROPOSED METHOD

The proposed system has semantic similarity, sentence similarity and good classifications results were achieved by using Wordnet based relations. The Term Matrix is constructed using the Wordnet lexical database.Retrieving relevant information with semantic relations by using hybrid approaches tools and techniques were studied and analyzed. Proposed system identified the NLTK (Natural Language Processing Tool Kit) performs well for the text classification. In this tools there are many built in algorithms embedded and which helps for good classification.The syntax and semantic oriented structure analysis can be done easily.

III.A. ARCHITECTURE DIAGRAM

The proposed architecture is as follow.

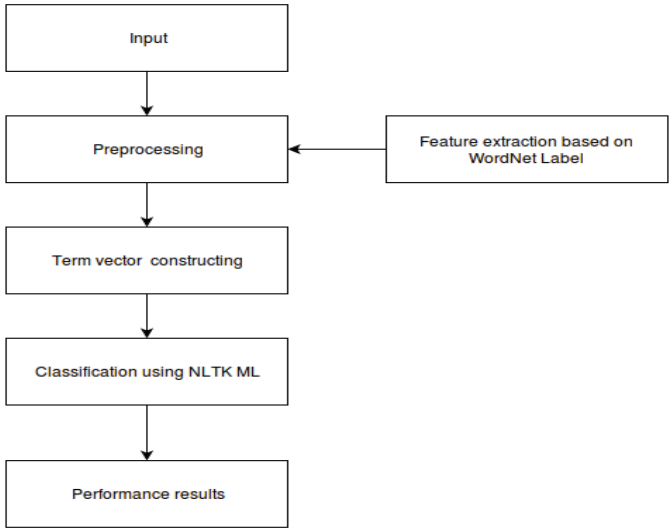

Fig 1. Proposed Architecture for Text Classification

The input is fed into data preprocessing to extract the features. Term Vector matrix is constructed based on Wordnet based relations. The Weights are added to Term Matrix based on similarity. If the words are more similar then weights are added to the corresponding term. It is very much useful for best classification results. Finally the performance results were taken and analysed for good pattern extraction.

III.B. Implementation steps

Pre-processing
Machine Learning/NLTK algorithm
Classification
Result

Pre-processing

In this module, data is getting pre-processed, removing stop words and stemming.

Machine Learning/NLTK algorithm

In this module Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Also it uses python NLTK, tokenizer divides a text into a list of sentences, by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences.

Classification

Detecting patterns is a central part of Natural Language Processing. Classification is the task of choosing the correct class label for a given input. In basic classification tasks, each input is considered in isolation from all other inputs, and the set of labels is defined in advance.

IV. RESULT AND DISCUSSIONS

The datasets used in this paper are downloaded from UCI Repository.This repository contains twenty newsgroups dataset for text analysis. In this paper ten newsgroups dataset are taken for the initial analysis. The

experiments are performed on ten newsgroups, they are Computer graphics, IBM PC Hardware, Autos, Baseball, Hockey, Electronics, Medical, Space, Politics and Religion.

Data Set for Experiments	Classes	Number of Documents	Classifier accuracy Euclideandistance		Classifier accuracyCosine distance	
			Previous	Proposed	Previous	Proposed
10 Newsgroup	Computer Graphics	486	90%	95.1%	89%	96.2%
	IBM PC Hardware	491				
	Autos	495				
	Baseball	497				
	Hockey	499				
	Electronics	490				
	Medical	495				
	Space	493				
	Politics	387				
	Religion	314				

Table 1: Dataset details

Results comparison Chart

The datasets were taken and results were analysed. Our proposed method increases the performance results by 95.1% and 96.2%. The performance results were shown in graph.

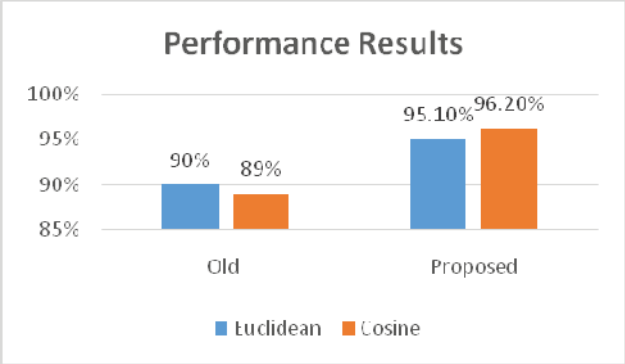

Fig 2: Results comparison

V. CONCLUSION AND FUTURE WORK

As per the analysis from existing references, all existing system has lagging in documents representations and semantic similarity. In this work the proposed system which has improved semantic similarity, sentence similarity and good way of classification. The K-Means classification is used for large sets of text documents. So that information retrieval can be performed faster and accurately. The precision and recall improved over 0.5% over other methods.

The future work may concentrate the more semantic relations to construct Term Matrix. Also the ontology based classification and construct ontology for the dataset produces good results.

REFERENCES

[1] Anil Kumar Patidar, Jitendra Agrawal, Nishchol Mishra, (2012), 'Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbour Clustering Approach', International Journal of Computer Applications (0975 – 8887) Volume 40– No.16.
[2] Anna Huang, (2008), 'Similarity Measures for Text Document Clustering', NZCSRSC, Christchurch, New Zealand.
[3] Aurangzeb Khan, BaharumBaharudin, Lam Hong Lee, Khairullah khan, (2010) 'A Review of Machine Learning Algorithms for Text-Documents Classification', Journal Of Advances In Information Technology, VOL.1, NO. 1.
[4] Pratiksha Y. Pawar,Gawande S. H, Member, IACSIT, (2012) 'A Comparative Study on Different Types of Approaches to Text Categorization', International Journal of Machine Learning and Computing, Vol. 2, No. 4.
[5] RedaSiblini, Leila Kosseim, (2013), 'Using a Weighted Semantic Network for Lexical Semantic Relatedness', Proceedings of Recent Advances in Natural Language Processing, pages 610–618, Hissar, Bulgaria.
[6] Simone Marinai+, Marco Gori, Giovanni Soda, 'Artificial Neural Networks for Document Analysis and Recognition'.

[7] VandanaKorde, NamrataMahender C, (2012) 'Text Classification and Classifiers: ASurvey', International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2
[8] VasileRus, MihaiLintean, RajendraBanjade, NobalNiraula,DanStefanescu, (2013), 'SEMILAR: The Semantic Similarity Toolkit', Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 163–168, Sofia, Bulgaria,Association for Computational Linguistics.
[9] Vijay S, (2012) 'A Combined Method to Measure the Semantic Similarity between Words', International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-ETIC2011.
[10] WeiweiGuo, Mona Diab, (2013), 'Improving Lexical Semantics for Sentential Semantics: ModelingSelectional Preference and SimilarWords in a Latent Variable Model', Proceedings of NAACL-HLT, pages 739–745, Atlanta, Georgia, 9–14, Association for Computational Linguistics.