International Journal of Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-2)

Research Article | **February 2017**

Improved Nearest Neighbour Approach for Document Categorization

Rimpy Wadhawan*
M.Tech. Research Scholar,
Galaxy Global Imperial Technical Campus,
Ambala, Haryana, India

Saurabh Mittal
Associate Professor,
Galaxy Global Imperial Technical Campus,
Ambala, Haryana, India

Abstract:

Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Text Categorization is an issue in Data mining. The application of Text clustering can be categorized to two types, online and offline. Online applications are usually constrained by efficiency problems when compared to offline applications.Text classification or Text categorization is a problem in library science, information science and computer science. Text categorization is used for sort the useful text and classifies the text by content. Text categorization is text classification. It is an approach of machine learning in the form of Natural Language Processing (NLP). The task is to assign a text to one or more classes or categories. This may be done "manually" or algorithmically. Texts may be classified according to their subjects or according to other attributes. In our research dataset is used and read the texts. The special symbols, stemming, and stop words are removed. Lowercase conversion performed to reduce the time. The occurrence of repeated words also measured. The tf-idf also calculated for vector space model. For the evaluation of performance precision, recall and f-measure also calculated.

Keywords: Data Mining, Text Mining, TF, IDF, TF-IDF

I. INTRODUCTION

Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

A. Text Mining

Text Mining [11] 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 with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text.

B. Categorization

Categorization involves identifying the main themes of a document by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often treat the document as a "bag of words." It does not attempt to process the actual information as information extraction does. Rather, categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on a thesaurus for which topics are predefined, and relationships are identified by looking for broad terms, narrower terms, synonyms, and related terms. Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic. As with summarization, categorization can be used with topic tracking to further specify the relevance of a document to a person seeking information on a topic. The documents returned from topic tracking could be ranked by content weights so that individuals could give priority to the most relevant documents first. Categorization can be used in a number of application domains. Many businesses and industries provide customer support or have to answer questions on a variety of topics from their customers. If they can use

categorization schemes to classify the documents by topic, then customers or endusers will be able to access the information they seek much more readily. The goal of text categorization is to classify a set of documents into a fixed number of predefined categories. Each document may belong to more than one class. Using supervised learning algorithms [13], the objective is to learn classifiers from known examples (labeled documents) and perform the classification automatically on unknown examples (unlabelled documents). Figure.1 shows the overall flow diagram of the text categorization task. Consider a set of labelled documents from a source D = [d1,d2,….dn] belonging to a set of classes C = [c1,c2,…,cp].

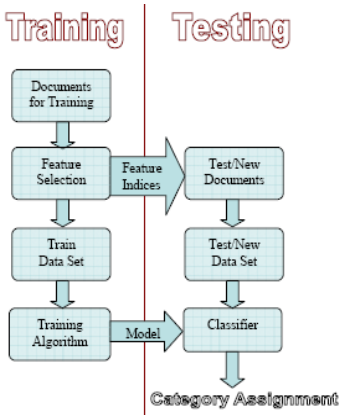

Fig. 1: Flow Diagram of Text Categorization

C. Text Categorization in Data Mining

We have been given a predefined [12] set of natural language text then the method of labelling natural language texts with reference to thematic division is called Text Categorization (TC).There was an extensive work on Text Categorization in early 60s but this field was evolved gradually, and in early 90s it has gained prominent status and has become a major sub field of the Computer and Information Systems Engineering discipline. Obviously there is a role of increased power of software applications and the high availability of more powerful hardware in the emergence of Text Categorization (TC). There are now different applications of Text Categorization (TC) in many contexts. Some of the applications are Controlled Vocabulary Based Document Indexing, Document Filtering, Automated Meta Data Generation, Word Sense Disambiguation and Population of Hierarchical Catalogues of Web Resources. Generally speaking, Text Categorization (TC) is now being applied in multiple contexts covering any application requiring document organization, selective document dispatching and adaptive document dispatching. Text Categorization (TC) can be applied to the data which is in the form of natural language text. The natural language text is divided or categorized among subset of texts and labelled according to the theme which is the main idea or subject. Text Categorization is applied on online newspapers, online news channels, e-papers, web search engines because these web technologies incorporate search and retrieval of data in the form of text.

II. RELATED WORK

Mohammed G. H. et. al. [1] presented an efficient rule-based method for categorizing free text documents. The contributions of this research are the formation of lexical syntactic patterns as basic classification features, a categorization framework that addresses the problem of classifying free text with minimal label description, and an efficient learning algorithm in terms of time complexity and F-measure.

Jiana Meng et. al. gives[2] proposes a two-stage feature selection algorithm. Firstly, they select features by the FCD feature selection method to reduce the feature numbers observably. Secondly, they apply LSI to construct a new conceptual vector space.

Lam Hong Lee et. al. [3] gives High Relevance Keyword Extraction (HRKE) facility is introduced to Bayesian text classification to perform feature/keyword extraction during the classifying stage, without needing extensive pre-classification processes.

M. Maharasi et. Al. [4] proposed the compactness of the appearances of a word and the position of the first appearance of a word are used. Three types of compactness-based features and the position-of-the-first-appearance-based features are implemented to reflect different considerations.

Kostas Fragos et. al. [5] proposed a method to improve performance in biomedical article classification. They use Naïve Bayes and Maximum Entropy classifiers to classify real world biomedical articles. They describe a technique based on chi-square measure to discard irrelevant information from the data and to identify the most relevant keywords to the classification task.

Shweta Taneja et. Al. [6] shown the major shortcomings affecting the traditional KNN algorithm and reviewed some improvements made to overcome them. Based on the analysis, they present our proposed KNN algorithm using dynamic selected, attribute weighted and distance weighted techniques. This proposed algorithm improves the accuracy of classification and reduces the execution time. It is a blend of classification and clustering techniques.

Deqing Wang et. Al. [7] proposed a t-test feature selection approach based on term frequency. The student t-test is used to assess whether the averaged term frequencies of a term between two classes are statistically different from each other by calculating a ratio between the difference of two class means and the variability of the two classes. Then they

compare our approach with the state-of-the-art methods on two common text corpora using three classifiers in terms of macro-F1 and micro-F1.

Anagha Kulkarni et. Al. [8] finds contexts of documents using pattern based clustering. Even though pattern based clustering is widely applied in gene expression or chromosome matching, it is not very common in text mining. The work proposed in this paper uses CSC to find contexts of clusters. The research reported in this paper suggests that CSC calculates closeness between patterns of documents.

"Mahoshadha" [9] project focused on retrieving the most accurate answers to the any given query. According to the test results the goal has accomplished with 98% of accuracy and with a high efficiency of generating the response. "Mahoshadha" opens new paths in QA systems.

Goyal Shubham [10] gives a methodology for the classification of sentiments was developed in this thesis for food price crisis in Indian market. Twitter API was used for streaming of tweets. The streamed tweets was filtered for relevant content and stored in a database. The several steps of pre-processing were applied on it and the tweets were removed from special characters, stop word, tokenized, etc. Stemming was done to all words in order to extract the root words.

III. PROPOSED ALGORITHM

A. Problem Definition

The concepts and techniques of text categorization is being used today in various fields. We also know that "Curse of Dimensionality" is the main problem in in text documents. A number of methods have been suggested for reducing the dimensionality of text documents. Some of them are: nonlinear dimension reduction techniques, discretizing high-dimensional data, latent semantic indexing (LSI) and Document Frequency (DF) etc. We have proposed a lexical approach, where we identify tokens or lexemes in the document. Each individual document is represented as a vector of tokens.

We use semantic similarity to find out the similar words which reduces the time where the idea of semantic similarity between words is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation.

B. Objectives

The previous algorithm was based on Weight based feature extraction strategy. Following are the objectives of iKNN algorithm that will be used in our algorithm:

The previous algorithm was based on Weight based feature extraction strategy. Following are the objectives of iKNN algorithm that will be used in our algorithm:

- Semantic similarity used for finding the meaningful words from two different words eg. lie and lies, the base word is lie, the meaning of both two words are same. So, uses the base word.
- In topic modelling, find out those terms that occur again and again in the document.
- Gain Ratio of threshold 0.01 will be used to improve the work.
- Improvement in different variant of KNN measures precision, recall and f-measure.

C. Proposed System Model

The proposed system can be summarized into three main steps that are integrated to give accurate results: text document representation, classifier construction and performance evaluation.

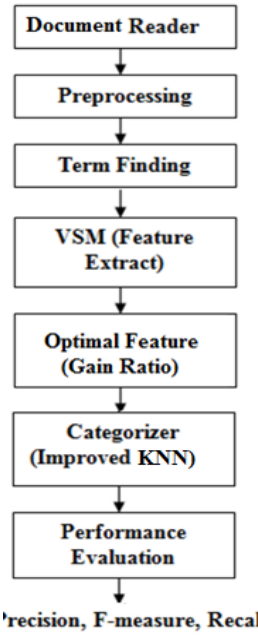

Fig. 2: The Proposed Text Categorization System Framework

D. Pseudo code of Proposed Algorithm

Step 1: Read the documents from which texts are to be categorized.
Step 2: Documents are scanned and pre-processed where stop words, special characters and stemming words (Improvement) are removed.
Step 3: Vector Space Model (VSM) is used to give a matrix of dimension n*m to find TF-IDF.
Step 4: Now, Optimal Feature is used where data having Gain Ratio (min throughput) of 0.1 are considered rest are discarded.
Step 5: Now extracted data is sent to classifier where KNN is used to classify the data.
Step 6: Finally, Performance is evaluated using f-measure, precision and recall.

E. Performance Evaluation

The performance evaluation uses the three methods to evaluate the result. The methods are f-measure, precision and recall. The formulas for precision, recall and f-measure are given below:

$$\text{precision} = \frac{|\{\text{relevent text}\} \cap \{\text{retrived text}\}|}{|\{\text{retrived text}\}|}$$

$$\text{recall} = \frac{|\{\text{relevent text}\} \cap \{\text{retrived texts}\}|}{|\{\text{relevent text}\}|}$$

$$\text{F} - \text{measure} = 2.\frac{\text{precision.recall}}{\text{precision} + \text{recall}}$$

F. Software Tool

We are using NetBeans which is a software development platform written in Java. The NetBeans Platform allows applications to be developed from a set of modular software components called modules. Applications based on the NetBeans Platform, including the NetBeans integrated development environment (IDE), can be extended by third party developers.

IV. RESULTS AND DISCUSSION

Our proposed work results are in the form of graph. It defines the result of precision, recall and f-measure. It is also show the time complexity of previous as well as proposed algorithm.

The results in the form of graph and description of results are shown below:

A. Time Complexity

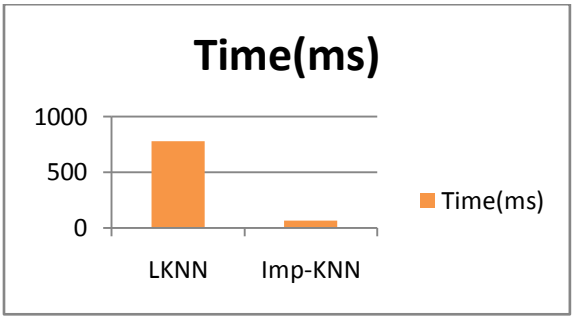

Fig. 3 Time complexity comparison between LKNN and iKNN

The figure 3 shows the time complexity of LKNN and iKNN. Time complexity in LKNN was 777 ms whereas in proposed algorithm it falls to 62 ms, which shows that proposed algorithm is fast in processing. So our proposed iKNN algorithm reduced the time complexity 125% than LKNN algorithm.

B. Precision

The figure 4 shows the normalized value of precision. In LKNN, we get the precision value as 0.819071 whereas in iKNN the value comes out to 0.985. The precision of iKNN is 20.00% (approx.) greater than LKNN. This increase in value gives a strong precision as compared to LKNN.

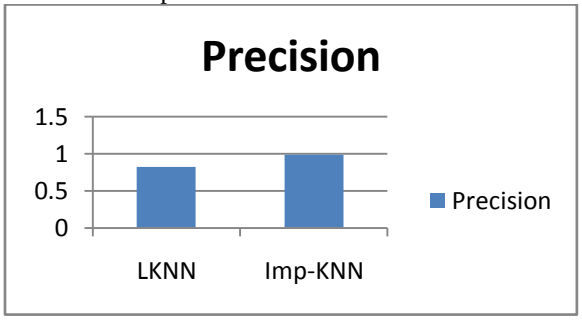

Fig. 4 Precision comparison between LKNN and iKNN

C. Recall

The figure 5 shows the normalized value of recall. LKNN gives the value of recall to 0.805, but proposed algorithm give this value as 1.0 which shows that the recall measure in proposed algorithm is exactly accurate. The recall of iKNN is 20.00% greater than LKNN.

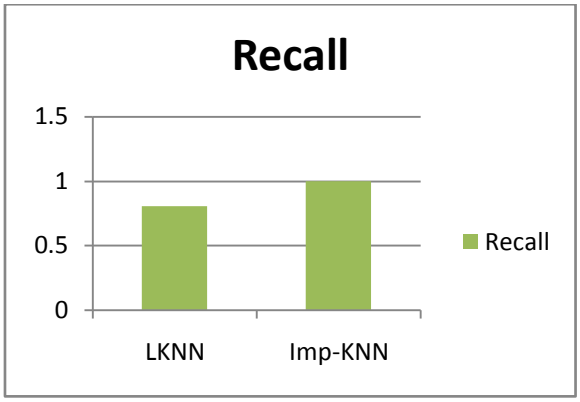

Fig. 5 Recall comparisons between LKNN and iKNN

D. F-measure

The figure 6 shows the result of LKNN and iKNN in terms of F-measure. The normalized value of f-measure shows in the following graph w.r.t. values of LKNN as 0.802826158 and that of proposed iKNN as 0.992443325. This shows the improved value of F-measure is almost equal to 1.0 . The f-measure of iKNN is 23.00% greater than LKNN.

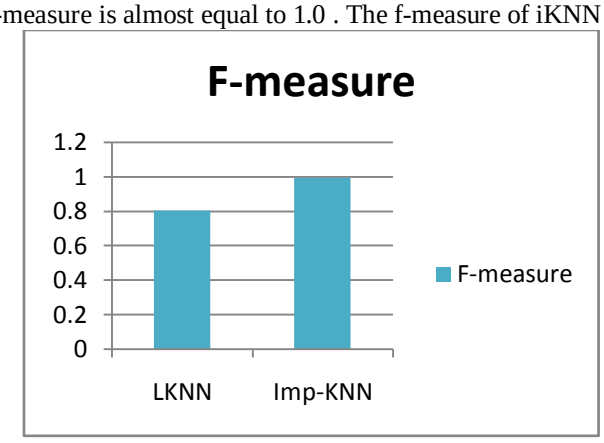

Figure 6 F-measure comparison between LKNN and iKNN

The overall improvement in result of Precision, Recall and F-measure gives that the proposed algorithm is more accurate than the previous one. Also the time complexity of iKNN is improved up to 125% as compared to LKNN.

V. CONCLUSION AND FUTURE WORK

Text categorization is used to divide the documents into different clusters. Same type of data is used in a document. Our goal is to reduce the time complexity and increases the accuracy. We performed text categorization on mini newsgroup dataset. We propose a new algorithm named as improved KNN (iKNN). The iKNN gives us better result than existing LKNN algorithm. The precision, recall and f-measure of improved KNN achieve better result than existing KNN algorithm. In existing algorithm there was no steaming of data. But we provide this steaming of data. Also optimal feature extraction is used in our work which was not in previous work. The result shows high accuracy in less time.

In our dissertation, we have been worked with single dataset. In future, iKNN algorithm can also used with multiple datasets. We can also use the iKNN algorithm with other classifier.

REFERENCES

[1] Mohammed G. H. Al Zamil , Aysu Betin Can, ROLEX-SP: Rules of lexical syntactic patterns for free text categorization, Knowledge-Based Systems, v.24 n.1, p.58-65, February, 2011 [doi>10.1016/j.knosys.2010.07.005].

[2] Jiana Meng, Hongfei Lin, Yuhai Yu, "A two-stage feature selection method for text categorization" in Computers and Mathematics with Applications 62 (2011), pp. no. 2793–2800.

[3] Lam Hong Lee , Dino Isa , Wou Onn Choo , Wen Yeen Chue, High Relevance Keyword Extraction facility for Bayesian text classification on different domains of varying characteristic, Expert Systems with Applications: An International Journal, v.39 n.1, p.1147-1155, January, 2012 [doi>10.1016/j.eswa.2011.07.116].

[4] M. Maharasi1, P. Jeyabharathi2, A. Sivasankari3, "Text Categorization Using First Appearance And Distribution Of Words", in Int. Journal of Engineering Research and Applications www.ijera.com Vol. 3, Issue 5, Sep-Oct 2013, pp.451-454.

[5] Kostas Fragos , Christos Skourlas, Toward Improving Classification of Real World Biomedical Articles, Proceedings of the 18th Panhellenic Conference on Informatics, October 02-04, 2014, Athens, Greece [doi>10.1145/2645791.2645848].

[6] Shweta Taneja, Charu Gupta, Kratika Goyal , Dharna Gureja, "An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering", in Fourth International Conference on Advanced Computing & Communication Technologies-2014,pp.-325-329.

[7] Deqing Wang, Hui Zhang, Rui Liu, Weifeng Lv , Datao Wan," t-Test feature selection approach based on term frequency for text categorization", in Pattern Recognition Letters 45 (2014) pp. no. 1–10.

[8] Anagha Kulkarni, Vrinda Tokekar, Parag Kulkarni, "Discovering Context of Labeled Text Documents using Context Similarity Coefficient" in Procedia Computer Science 49 (2015) 118 – 127.

[9] J. A. T. K. Jayakody, T. S. K. Gamlath, W. A. N. Lasantha , K. M. K. P. Premachandra, A. Nugaliyadde, Y. Mallawarachchi, ""Mahoshadha", The Sinhala Tagged Corpus based Question Answering System", in Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1, July-2016.

[10] Goyal Shubham, International Journal of Advance Research, Ideas and Innovations in Technology, ISSN: 2454-132X, Volume2, Issue5, pp. no. 1-9, 2016.

[11] Berry Michael W., (2004), "Automatic Discovery ofSimilar Words", in "Survey of Text Mining: Clustering,Classification and Retrieval", Springer Verlag, New York.

[12] Ahmed Faraz"An Elaboration Of Text Categorization And Automatic Text Classification Through Mathematical And Graphical Modelling" in Computer Science & Engineering: An International Journal (CSEIJ), Vol.5, No.2/3, June 2015.

[13] Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati, "Experiments on Supervised Learning Algorithms for Text Categorization", International Conference , IEEE computer society, 1-8, (2005).

