A Novel Approach for Keyword Extraction in learning objects using Text Mining

Akshita Thakkar[1],Sandeep Chauhan[2]

[1] Student ME, Computer Science & Engineering, KIRC-Kalol, Gujarat,India
[2] Assistant Professor, Computer Science & Engineering, KIRC-Kalol, Gujarat,India

ABSTRACT

This paper address the problem of extracting learning objects or keywords from bunch of Documents ,with the goal of use this objects for various purpose like Resume filtering ,Email filtering ,content classification etc. Keyword extraction, concept finding are in learning objects is very important subject in today's eLearning environment. Keywords are subset of words that contains the useful information about the content of the document. Keyword extraction is a process that is used to get the important keywords from documents. In this proposed System I Calculate the TF-IDF of each word, then Decision tree algorithm is used for feature selection process using wordnet dictionary. WordNet is a lexical database of English which is used to find similarity from the candidate words. The words having highest similarity are taken as keywords.

Keyword – *Keyword Extraction, Text Mining, TF-IDF, Decision Tree,Wordnet*

1. Introduction

With the advancement of technology, more and more data is available in digital form. Among which, most of the data (approx. 85%) is in unstructured textual form. [8].The vivid increase of documents demands effectual retrieval and organizing methods mainly for large documents. So there is a need for automatically retrieval of useful information from the large amount of textual data. Manual analysis of the unstructured textual information is impractical, and as a result, text mining techniques are being developed to convert the unstructured information to structured format.[7] Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases , but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. Using Text mining ,Readers benefit from keywords because they can judge more quickly whether the text is worth reading. Website creators benefit from keywords because they can group similar content by its topics. Algorithm programmers benefit from keywords because they reduce the dimensionality of text to the most important features.[9]

2. Literature Review

Paper presented by Maryam Habibi and Andrei Popescu-Belis [1],.they proposed a novel diverse keyword extraction technique which covers the maximal number of important topics in a fragment. Then, to reduce the noisy effect on queries of the mixture of topics in a keyword set, they proposed a clustering technique to divide the set of keywords into smaller topically-independent subsets constituting implicit queries. they built single queries from the keyword sets provided by the D(.75), TS and WF keyword extraction methods,and compared the three resulting document sets. Secondly, they built multiple queries from the same methods and performed similar comparisons between the resulting document sets. Finally,they compared the best results of multiple queries with the best results of single queries Their experiments showed that the diverse keyword extraction method provides on average the most representative keyword sets,.

Paper presented by Ahmad A. Kardan, Farzad Farahmandnia, Amin omidvar [2] that A novel approach is proposed to improve keyword extraction in learning objects. This proposed method use standard algorithms of text mining to extract keywords from learning objects. After that using WordNet dictionary, concept distances between keywords are calculated. In final step, keywords with highest similarity will be selected as output keywords.

Paper presented by Menaka S, Radha N [3] that proposed system Documents are collected from different journals keywords are extracted from documents using TF-IDF and WordNet. TF-IDF algorithm is used to select the appropriate candidate words. WordNet is a lexical database of English that used to find similarity between the candidate words. The words having highest similarity are taken as keywords. These keywords are stored in the database for classification. The 10 fold cross validation is used to evaluate the robustness of the classifiers. The training time and the prediction accuracy are two conditions used to evaluate the prediction accuracy of the each model and the performances of the trained models is compared. The experiment has been done using Decision tree, K-Nearest Neighbour (KNN) and Naive Bayes algorithms and its performance are analysed. From the result, accuracy for text classification is better using Decision tree algorithm compared to other algorithms.
Paper presented by Jadhav Bhushan , Warke Pushkar , Kuchekar Shivaji , Kadam Nikhil [4] they take research papers as documents ,as we spend 2 or more hours to read a single paper. So it is helpful to provide better verification of the document (research paper). This architecture works on Distributed Knowledge Database System related with database, operating system and topic of programming. It works on the specific keyword based on text mining technique initially. It will search the base keyword of the content from the knowledge database. Proposed work uses the search engine based on clustering and text mining. K means algorithm is used to establish a cluster, which permit us to identify approximate text to search using predefined patterns. It will be used to send the parameters for classification of documents (research papers) in this case the search engine will use five clusters to achieve implementation.
Paper Presented by Dr. Shilpa Dang1, Peerzada Hamid Ahmad [7] that is review of all Text mining techniques with respect to different application areas. Also compared the text mining technique with respect to tools, algorithm used, characteristic and models.

3. Background Theory

3.1 Keyword extraction

Keywords can be considered as important words of documents and short forms or base word of their summaries. Keyword extraction is technique for number of text mining related tasks like webpage retrieval, summarization, and document clustering and document retrieval. The main aim of keyword extraction is to extract more relevance from the text. First step is to select the desired documents like text files or pdf files. And it can be pre-processed.

***Stop Word Remove*:** The Stop words are a part of natural language which do not have meaning in a retrieval system. The reason that stop-words can be removed from a text is that they make the text look complex, less important and heavier for analysts. Removing stop words reduces the dimensionality of term space. Prepositions, articles, and pro-nouns etc. don't have meaning in documents, so these are treated as Stop word. For Example in, the, with, an, a, etc. Stop words are removed from documents because those words are not considered as keywords in text mining applications.

Stemming: A Stemming techniques are used to find out the stem/ root of a word. Stemming converts words to their roots which incorporates a great deal of language-dependent linguistic knowledge. For example, the words, used by, using, used, all can be stemmed to the word 'use'. In the current work, the Porter Stemmer algorithm which is the most common algorithm that is in used.

N-Grams Technique: N-grams of texts are extensively used in text mining and natural language processing tasks. They are basically a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward (although you can move X words forward in more advanced scenarios).

Term Frequency-Inverse Document Frequency: Term Frequency–Inverse Document Frequency (tf-idf) is a numerical statistic which reveals that a word is how important to a document in a collection. Tf-idf is often used as a weighting factor in information retrieval and text mining.

- ✓ TF= no .of term frequency/ total no of words in a doc
- ✓ IDF no.of docs with the words / total no of docs
- ✓ Tf-idf=tf*idf

Feature Selection: Feature selection is a process commonly used in Machine Learning field to reduce the dimensionality of the feature space. The subset of the features available in the data is keywords are selected out.

Decision tree: The decision tree rebuilds the manual categorization of documents by constructing well- defined tree structure. In decision tree structure, leaves represent the category of documents and branches represent conjunctions of features that lead to those categories. The well-structured decision tree can easily classify a document by putting it as the root node of the tree and let it run until it reaches a certain leaf through the query structure. Leaf, which represents the goal for the classification of the document.

To generate a decision tree, the ID3 algorithm needs per branch at most as many decisions as features are given.

a. no backtracking takes place
b. local optimization of decision trees

4. Proposed Work

Step 1: Read Documenets

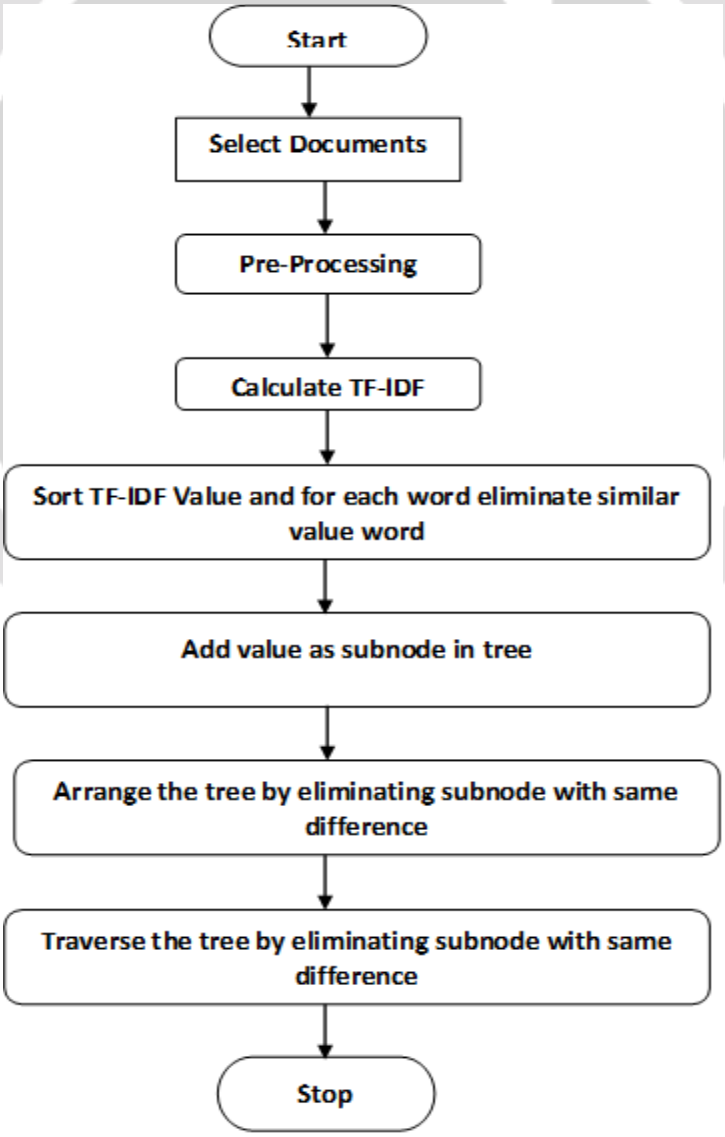

Fig -1 proposed work flow chart

Step 2: In preprocessing Stopping and stemming is performed. After this tokens are generated using n-grams technique.
Step 3: For each compare with WordNet semantics and eliminate for similar meaning word.
Step 4: Build a tree with new words and if a similar word meaning found then add as a sub node.
Step 5: Calculate the difference between sub nodes and arrange it in descending order.
Step 6: Take keywords which are most relevant display them.

5. RESULT

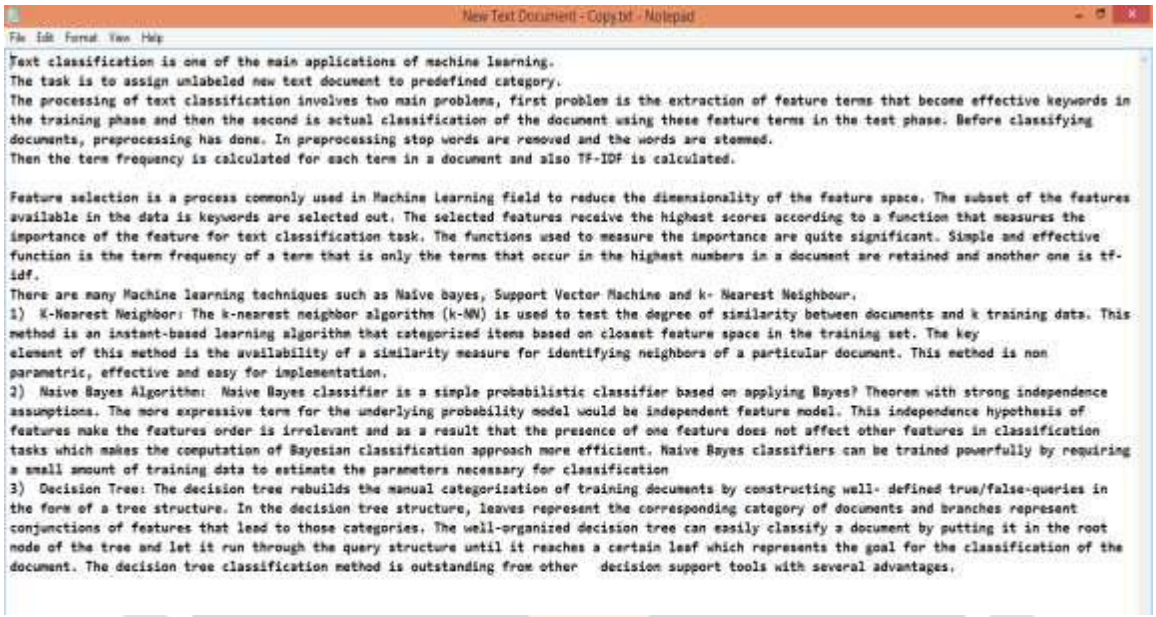

Fig- 2 Document as Input

Fig-3 Keywords as output

5. CONCLUSION

In this paper a novel approach is proposed to improve and increase keyword extraction accuracy in learning objects. From research decision tree algorithm has better accuracy for text classification compared to other algorithms [3].So in proposed method I use decision tree algorithm for feature selection using wordnet dictionary. This approach use Porter method for stemming, Then, using WordNet dictionary, concept distances between keywords are calculated. In final step, keywords having highest similarity will be selected as output keywords.

6. REFERENCES

PAPERS

[1] Maryam Habibi and Andrei Popescu-Belis, "Keyword Extraction and Clustering for DocumentRecommendation in Conversations", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, APRIL 2015
[2] Ahmad A .Kardana, Farzad Farahmandniab, Amin Omidvarc, AWERProcedia Information Technology & Computer Science, "A novel approach for keyword extraction in learning objects using text mining and WordNet ", 2012
[3] Menaka S* and Radha N, "Text Classification using Keyword Extraction Technique", Volume 3, Issue 12, December 2013
[4] Jadhav Bhushan G, Warke Pushkar U, Kuchekar Shivaji P, Kadam Nikhil V , "Searching Research Papers Using Clustering and Text Mining",International Journal of Emerging Technology and Advanced Engineering.Volume 4, Issue 4, April 2014
[5] Zakaria Elberrichi, Abdelattif Rahmoun and Mohamed Amine Bentaalah,"Using WordNet for Text Categorization", The International Arab Journal of Information Technology, Vol. 5, No. 1, January 2008
[6] Ms. Anjali Ganesh Jivani , "A Comparative Study of Stemming Algorithms", IJCTA , NOV-DEC 2011
[7] Dr. Shilpa Dang, Peerzada Hamid Ahmad, "Review of Text Mining Techniques Associated with Various Application Areas", International Journal of Science and Research, Volume 4 Issue 2, February 2015
[8] Lokesh Kumar, Parul Kalra Bhatia, "Text Mining: Concepts, Process And Applications", Journal of Global Research in Computer Science, Volume 4, No. 3, March 2013

WEBSITE

[1] Alyona Medelyan, "NLP keyword extraction tutorial with RAKE and Maui", https://www.airpair.com/nlp/keyword-extraction-tutorial
[2] https://en.wikipedia.org/wiki/Stop_words