International Journal of Engineering & Technology, 7 (4.34) (2018) 437-440

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Document Categorization Using Decision Tree: Preliminary Study

Wan M.U. Noormanshah[1], Puteri N.E. Nohuddin[1]*, Zuraini Zainol[2]

[1]Institute of Visual Informatics, National University of Malaysia 43600 Bangi, Selangor, Malaysia
[2]Department of Computer Science, Faculty of Science and Defence Technology, Universiti Pertahanan Nasional Malaysia, Sungai Besi Camp, 57000 Kuala Lumpur, Malaysia
**Corresponding author E-mail: puteri.ivi@ukm.edu.my*

Abstract

This preliminaries study aims to propose a good classification technique that capable of doing document classification based on text mining technique and create an algorithm to automatically classify document according to its folder based on document's content while able to do sentiment analyses to data sets and summarize it. The objective of this paper to identify an efficient text mining classification technique which can resulted with highest accuracy of classifying document into document folder, capable of extracting valuable information from context-based term that can be used as an output for algorithm to do automatic classification and evaluate the classification technique. Methodology of this study comprises in 5 modules which is 1) Document collection, 2) Pre-Processing Stage, 3) Term Frequency-Inversed Document Frequency, 4) Classification Technique and Algorithm, and lastly 5) Evaluation and Visualization of the classification result. The proposed framework will have utilized Term Frequency-Inversed Document Frequency (TF-IDF) and Decision Tree technique which TF-IDF used as purposes to rank all the terms based on most frequent to least frequent terms so, while decision tree function as decision making in terms of deciding which folder the document belongs to.

Keywords: *Data mining; Unstructured data; Decision tree; Text mining; Text frequency – inversed document frequency.*

1. Introduction

We lived in an era that computing technology grows so fast and data collecting becomes notable and contribute too many fields of work such as in medical used, business, education, reference, report, etc. Real world data have many type which is qualitative, quantitative, discrete, etc. These data can be recorded and visualized in variety of mediums such as electronic document, and databases. With mountains amount of data, process of data keeping is needed in terms of organize, classify, analyses, and storing the data will be vital. Well organized data handling will result in smooth daily operation of an organization and will able to extract needed information from dumping data. Data handling used in many field for example data mining, knowledge discovery, pattern recognition, etc. Data mining also known as knowledge discovery in databases is the process of extracting hidden useful knowledge through large data set with help of tools to analyses data. Classification is one of data mining components that used to analyse and result in predict set of data according to its target class that a data belongs to. The intention of this research is to suggest a classification technique that able to analyse document content and auto assign the document according to its folder using text mining technique. The rest of this paper is arranged as follows. Section 2 will be discussing on Text Mining, Decision Tree and Term Frequency- Inversed Document Frequency. Section 3 briefly explained the detail of the proposed framework for analysing and classifying document. Lastly, we end this paper with future work in conclusion section.

2. Related Work

Section 2 consist of 3 sub section that explained each related work in this preliminaries study. First sub-section is about introduction of text mining, followed by term frequency inversed document frequency and end with comparison of classification technique in text mining.

2.1. Text Mining

Text Mining (TM) is one of an analytics process, it was formulated to execute a task in analyzing a collection of unstructured textual materials in deriving high-quality information and essential knowledge covered by raw texts and TM specified in takes care of unstructured information. TM is one of branch in data mining, which requires numerous disciplines areas such as Information Retrieval (IR), Computational Linguistics, Natural Language Processing (NLP) Web Mining and Statistics. [1].
Since TM is related to data mining, the method of withdraw potential valuable information should basically go through the same fundamental procedure of data mining. TM is explicitly used to separate high-quality info in a domain of text and it mines information and knowledge from a mountain of text. Additionally, with a combination of TM and Term Document Matrix (TDM) [2] its competent to indexed and count all terms appear in each document in table form which arranged by column for terms appeared in a document and row represents the document identification or vice versa.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2.2. Term Frequency Inversed Document Frequency

TF-IDF is partially one of TM technique that utilized for document classification. TF-IDF is an information retrieval technique that specifies and measure the weight of each term frequency (TF) and its inversed document frequency (IDF) [3]. This technique frequently appears in NLP or IR [4]. This technique is utilized to calculate each term appears which evaluates the significant of terms in a document collection. The significant value of the term is rise proportionally to the number showed up in the documents [5]. Regularly, TF-IDF weight is composed by two components. First, TF is responsible for quantifying the normalized term frequency. Second, IDF is calculated according to the formula which is the total number of documents in a corpus is divided by the number of documents where the specific terms appeared, and the result will be multiplies with logarithm. The prescription of TF-IDF can be illustrated below:

- TF = (Total frequent term 'x' appears in a document) / (Sum of terms in the document)
- IDF = log e (Exact total of document / Total of documents with term 'x' in it)

Amid processing TF, all term can be considered as similarly vital in weight. Be that as it may, in some cases there is plenty of terms show up repeatedly in a document content that can be considered as less essential, for example, 'a', 'is', and 'of'. Thus, IDF computing helps user to load down the regular terms while scale up the rare ones that have important meaning to user. TF-IDF gives notable effect in information retrieval to rank term's results. Along these lines, in this research, we suggest TF-IDF as reduction attribute technique to be combined with decision tree as an absolutely factual method to assess the significance of words dependent on its frequency of occurrence in the document and in its related corpus.

2.3. Text Classification Technique and Comparison

Many studies have been conducted on building and finding most efficient classification technique for classify document. But the effectiveness of a classification technique is depending on attribute and environment of data set that applied into the technique. There is no most accurate or perfect technique to classify, but by following rules and attributes that have been set into each classification technique ones able to build an efficient and accurate classification algorithm based on suitable framework and environment of data set. Categorizing set of documents can be illustrates with help of classification technique as task that automatically classify document into its target class based on document content [6]. In data mining, several of classification technique can be found, for example K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Decision Tree (DT).

2.3.1. K-Nearest Neighbour (KNN)

KNN is one of the simplest algorithm for classification technique which is direct classifier, which data set are classified based on dependent on the class of the nearest neighbor [7]. Even though with such simplicity, its capable of giving high competitive results in terms of accuracy classified data. Basically, KNN operates by the majority vote of its neighbor and the object will be allocated to the most related among its k closest neighbor. KNN algorithm otherwise called:

- cases-based reasoning
- example-based reasoning
- instance-based learning
- memory-based reasoning
- lazy learning

KNN tends to perform very well on large number of data set. It is naturally non-linear and able to distinguish either the data set is linear or non-linear appropriated data. Lazy learning in KNN is not defines because of its apparent simplicity, but due to its does not learn a discriminative function from the training data, instead of that KNN memories the training data set.

2.3.2. Support Vector Machine (SVM)

SVM is supervised machine learning algorithm which capable to be used for regression and classification, but commonly used for classification issues. It's good to be applied in multidomain application in a big data environment [8]. SVM can be used in both linear and non-linear way with the use of a Kernel, SVM is excellent when one had restricted set of points in numerous measurements since it capable to discover linear separation that exists. In addition, SVM is universal learned [9]. In its basic form, SVM learns linear threshold function, however by easy plugin appropriate kernel their function can be level up to another future function. Nonetheless, SVM is mathematically complicated and computationally exorbitant. Plus, as indicated in [10], they concluded that KNN classification approach is more precise than SVM classification technique

2.3.3. Decision Tree (DT)

DT is one of classification technique in data mining that uses branches method to depict each feasible result of a decision making in each possible outcome [11]. DT comprises three kinds of node that frame an established a tree which a tree required to have 'root node', 'internal node', and 'leaf' [12]. DT breaks down a set of data into smaller and smaller subsets while at the same time an associated decision tress is incrementally build. Root node known as initial attribute or the topmost decision node in a tree which corresponds to the best predicator for a tree to make decision making that have zero incoming and outgoing edges. While, internal nodes have both incoming and outgoing edges at least one. Followed by leaf node which has no outgoing edges represents a classification or decision.

DT learn from data to approximate a sine curve with a set of IF-THEN rules and used for decision making. The deeper the tree, the more complex DT can be in decision rules and the fitter the model. Also, based on [13] the complexity of a tree will tends to affect the result of accuracy for a tree to do decision making. According to [14] DT much more convenient to do classification when it involved decision making, instead of able to compute both categorical and numerical data, it easily accessible and interpreted, involved less calculation, capable to illustrates relationship between dependent and independent variables and computationally low end. For document auto classification, DT is suitable to be applied into a simple framework that setting a set of rules and used for decision making to classify document based on its content into its category. Followed is a quick comparison table for all selected classification technique that discussed previously in this study.

Table 1: Comparison of Classification Techniques between KNN, SVM and Decision

KNN	SVM	DT
Can be used for continuous value inputs.	Can be used for continuous value inputs.	Can be used for continuous and categorical inputs.
Algorithm is simple with straight forward classifier easy understand.	Mathematically complex and hard to build own algorithm.	Data classification with less calculation involved. Easily understand.
It is automatically non-linear, able to detect linear and non-linear distributed data. Perform very well a lot of	Can be used in linear and non-linear ways and good with limited set of points in many dimensions.	It is non-linear classifier. Able to illustrate relationship between independent and dependent variables.

data points.		
Do classification by determine neighbour-hoods.	Do classification by searches for closest points.	Do classification by form a tree.
Computationally ex-pensive.	Computationally expensive to train	Computationally low end.
Not suitable for auto classification.	Suitable for auto classification tech-nique but a bit com-plex.	Suitable for auto classification tech-nique and less com-plex.
Time consuming.	Time consuming when processing large amount	Time consuming if involves multiple branches.

3. Framework of Keyword-Based Text Classification

In the comparison table can be seen that suitable technique to be used for auto-classification for document is DT. In comparison stated that DT is less complex in computing, reduce time consume, computationally low-end and much plainer during doing class label compared to SVM and KNN. In this section, proposed framework will be briefly discussed started from data collecting, pre-processing, analysis, and classification technique that will be utilized on Rstudio platform.

3.1. Proposed Framework

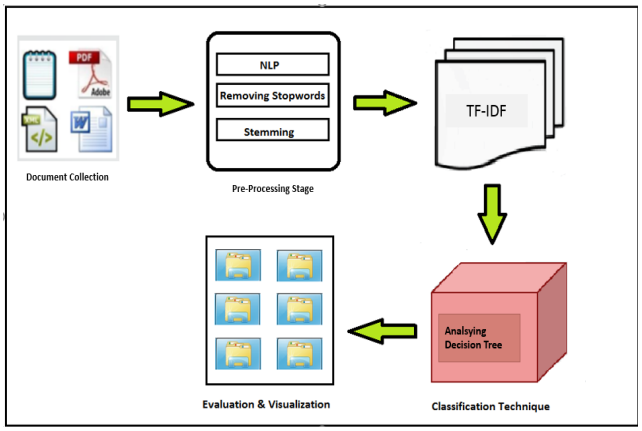

Fig. 1: Framework for Keyword-Based Text Classification (KBTC)

KBTC structure (see Figure 1) comprises of 5 modules which is (i) Document collection where all the collected document will be collected online from BBC online dataset. Next, (ii) Pre-Processing stage, which in this phase all raw document will be cleaned from unstructured text, noise and inconsistent data. All stopwords that appears will be remove so that process of extracting information and analyse document content based on terms appeared in document will be much easier and it can improve the quality of data. Also, in this process involved removing insignificant words and symbols. Stopwords are list of natural language words which have a little meaning or does not carry any meaning and information such as 'and', 'the', 'after', etc. Plus, stop words is known as irrelevant word for searching purposes because it occurs frequently. Stemming is process in reducing words into its root. For example, *Data mining is one of the process of analysing and exploring large amount of unstructured text data.* Pre-processing stage works for stopwords and stemming shown below in Table 2.

Table 2: Pre-processing stage

Function	Rscripts
Removing stop-words	Data mining one process analysing exploring large amount unstructured text data.
Stemming Pro-cess	Text mine process analyse explore large amount struc-ture text data

As shown in the Table 2, the irrelevant words of the text have been detached and cleansed through removing stop words process. Followed by stemming process that simplified the example string of text into its root words and produce output of significant word that facilitated the process analysing text document. Added that in this phase, stemming process is required so that terms that have same root terms can be prune into its basic root term, this process is required later for terms counting in TF-IDF and to facilitate the text analysis.

In this study, classification technique, DT will be selected to classify documents according to its folder. Followed by (iii) TF-IDF, the used of TF-IDF as explained before in section 2 of this preliminaries study which is for word or terms count, in KBTC it will be utilized for terms counting for each term that appeared in document collection and rank from most frequent terms to the least frequent. All cleaned document will be go through TF-DIF phase for terms counting and rank all the terms appear in each document started from most frequent term appears to the least. Follows by (iv) Classification technique which categorizing terms using DT technique. Before we can classify document and detect which folder the document belongs to, we need to analyse the correlation between each term against document. So that the result of the analysis can be utilized to be remark and tag as the category of document covered. Only then, classification can be done using DT technique. Which each document will go through each nodes of DT and tested either the document reaches the specification or attribute that belongs to tagged folder. Finally, (v) evaluation and visualization, algorithm for auto classification will used the result of classification to automatically send the documents into their folder.

3.2. RSTUDIO

Rstudio is an open source and extension of basic R statistical computing environment with integrated development environment (IDE). It is written in C++ and equipped with Qt framework for its graphical programming which result in organized interface instead of less graphical and basic command line script. Rstudio can be run either locally in desktop version or online through browser that connected to Rstudio server. Basically, Rstudio is tools that capable to do calculation, statistical computing, graphical and easy to generate raster-based model. The fascinating part of Rstudio is empowers user to create and control its own object, download existed package or create new package, and create function. Rstudio comes as a user-friendly tool that comes with integrated development environment that allows users to interact with R more readily. It is equipped with drop down menus, windows with multiple tabs, and many customization options. Rstudio by default when opened, users will found three windows that which is console window (location where commands are entered and the output is printed), environment and history windows (interactive list of loaded R object and list of key strokes entered in the consoles), and files,plot, packages, help window (which facilitate with files explorer, list of installed package, output location for help commands and help search window). Lastly the forth window which is source window can be found by user select a drop-down list and open new file for new Rscript.

4. Expected Outcomes

Expected results of this research is applicable in managing set of documents in terms of storing document, reduce the time needed to arrange or retrieve document according to its folder, carry out task for auto classification and replaced the old way practice in organized document with effortless system that resulted in increasing work efficiency with less cost. Plus, expected that KBTC framework able to completely work, reach the requirement and objective of this study also user friendly in term of document

management system application. Lastly, the present algorithm competent to do sentiment analysis of each document.

5. Conclusion

The purpose of this preliminaries study is to build an algorithm of classification and analysis of documents content in an effective way and accurately predict the document belongs to its category folder. Classification techniques in data mining was chosen based on analysis and comparison that suit to be apply in the proposed framework. The aftereffect of this research can be utilized as an enhancement in business operation, improve old way of inventory system, organizing database, categorized data, etc. Next future work, the proposed algorithm will be interpreted into real experimental, visualization and evaluation stage. All the end code algorithm will be built in Rstudio and the framework will be upgraded into tools that capable work in operating system.

References

[1] Inzalkar M, J Sharma. A survey on text mining-techniques and application. International Journal of Research in Science and Engineering, 2015, 24: 1–14.
[2] Zainol Z, Jaymes MTH, Nohuddin PNE. VisualUrText : A text analytics tool for unstructured textual data. Journal of Physics: Conference Series, 2018, 1018(1): 1-8.
[3] Trstenjak B, Mikac S, Donko D. KNN with TF-IDF based framework for text categorization. Procedia Engineering, 2014, 69: 1356–1364.
[4] Masuda K, Matsuzaki T, Tsujii J. Semantic search based on the online integration of NLP techniques. Procedia - Social and Behavioral Sciences, 2011, 27: 281–290.
[5] Zhang W, Yoshida T, Tang X. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Systems with Applications, 2011, 38(3): 2758–2765.
[6] Bijalwan V, Kumar V, Kumari P, Pascual J. KNN based machine learning approach for text and document mining. International Journal of Database Theory and Application, 2014, 7(1): 61–70.
[7] Moudden I El, Jouhari H. Learned model for human activity recognition based on dimensionality reduction. Proceedings of the Smart Application and Data Analysis for Smart Cities, 2014, pp. 1-6.
[8] Jabbar MA, Deekshatulu BL, Chandra P. Heart disease classification using nearest neighbor classifier with feature subset selection. Anale Seria Informatica, 2013, 11, 47-54.
[9] Joachims T. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the European Conference on Machine Learning, 1998, pp. 137-142.
[10] Manekar V, Waghmare K. Improving accuracy of SVM using hybrid cultural algorithm. International Journal of Computer Technology and Applications, 2014, 5(3): 1194–1197.
[11] Tan PN, Steinbach M, Kumar, V. Classification: Basic concepts, decision trees, and model evaluation. Introduction to Data Mining, 2006, 67(17): 145–205.
[12] Rokach L, Maimon O. Decision trees. In O. Maimon, & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook. Massachusetts: Springer, 2009, pp. 149–174.
[13] Lee J. A new approach of top-down induction of decision trees for knowledge discovery. PhD thesis, Iowa State University, 2008.
[14] Farooqui MA, Sheetlani J. Different classification technique for data mining in insurance industry using Weka, 2017, 19(1): 11–18.