A Review Of Document Categorization Techniques

[1]SHAIFALI GUPTA, [2]REENA RANI
[1]M.Tech Student ,Deptt. Of CSE
[2]Assistant professor Deptt. Of CSE
JMIT,Radaur,kurukshetra

ABSTRACT

Text categorization task have gained attention of researchers in last 10 years with the increase in web-based contents of documents. For seeking a particular document from internet or any large document collection text or document categorization is most helpful task. We demand some better system and improved machine learning classifiers to accomplish task of document categorization. Generally, traditional classification algorithms from machine learning field are used in text classification. These algorithms are mainly designed for structured data. In this paper the topic under discussion is about document categorization and its techniques i.e. NB, SVM, KNN etc.

Keywords— *Machine learning, Document categorization, Naïve Bayes Classifier (NB),Support Vector Machine (SVM), k-nearest neighbor (k-NN).*

I. INTRODUCTION

Data mining is the interdisciplinary subfield in computer science. It is the computational procedure of finding patterns in big data sets including methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall objective of the data mining process is to extract information from data set and change it into an understandable structure for further utilize. Beside the raw analysis step, it involves database and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is analysis step of the "knowledge discovery in databases" process, or KDD. Document categorization is a problem in which task is to assign a document to one or more classes or classifications. It can be done manually (or intellectually) and algorithmically. The intellectual classification of documents has mostly been the part of library science, while the algorithmic classification of documents is mainly in information science and computer science.

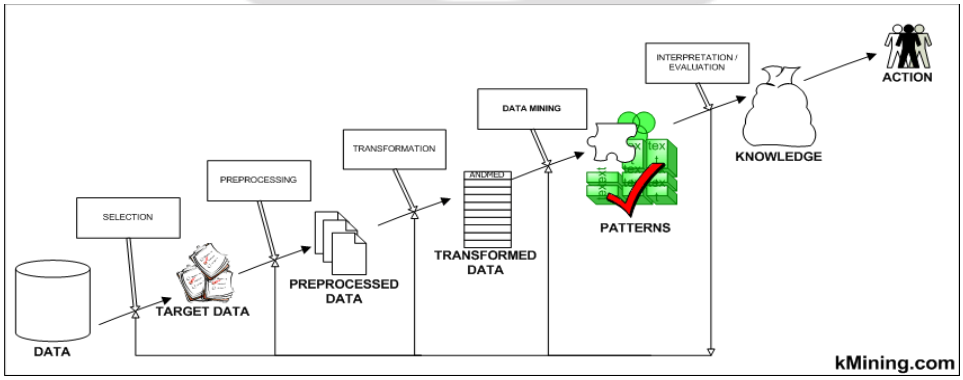

Fig 1 Data mining process

II. APPLICATIONS OF DOCUMENT CATEGORIZATION

- Spam filtering, a process which tries to observe e-mail spam messages from legitimate emails.
- Email routing, sending an mail sent to a general address to a specific address or mailbox depending on topic.
- Language identification, automatically deciding the language of a text.
- Genre classification, automatically determining the genre of a text.
- Readability assessment, automatically deciding the degree of readability of a text, either to find suitable materials for different age groups or reader types or as a part of a larger text simplification system.
- Sentiment analysis, deciding the attitude of a speaker and a writer with respect to some topic or the overall contextual polarity of a document.

III. TECHNIQUES OF DOCUMENT CATEGORIZATION

Automatic document classification techniques includes:

- Expectation maximization (EM)
- Naive Bayes classifier
- tf–idf
- Instantaneously trained neural networks
- Latent semantic indexing
- Support vector machines (SVM)
- Artificial neural network
- k-nearest neighbour algorithms
- Decision trees such as ID3 or C4.5
- Concept Mining
- Rough set-based classifier
- Soft set-based classifier
- Multiple-instance learning
- Natural language processing approaches

IV. PERFORMANCE MEASURES

In case of the term frequency tf(t,d), the simplest decision is to use raw frequency of a term in a document, i.e. the number of times the term t occurs in the document d. If we denote raw frequency of t by $f_{t,d}$, then the simple tf scheme is $\mathrm{tf}(t,d) = f_{t,d}$. Other possibilities include:

- Boolean "frequencies": tf(t,d) = 1 if t occurs in d and 0 else;
- Logarithmically scaled frequency: $\mathrm{tf}(t,d) = 1 + \log f_{t,d}$, or zero if $f_{t,d}$ is zero;
- Augmented frequency, to prevent a bias towards larger documents, e.g. raw frequency divided by the maximum raw frequency of any term in document:

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

The inverse document frequency is measure of how much information the word provides, that is, whether the term is common or rare across all the documents. It is the logarithmically scaled inverse fraction of the documents that contain word, obtained by dividing total number of documents by number of documents containing the term, and taking the logarithm of that quotient.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

with

- NN: total number of document in the corpus N = |D|
- $|\{d \in D : t \in d\}|$: number of documents in which the term "t" appear. If the term is not in corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to $1 + |\{d \in D : t \in d\}|$.

V. RELATED WORK

Murat Can Ganiz et all (2015) [1] Text classification is one of the key methods used in text mining. Basically, traditional classification algorithms from machine learning field are used in text classification. These algorithms are primarily designed for structured data. In this paper, we propose a new classifier for textual data, called Supervised Meaning Classifier (SMC). The new SMC classifier uses meaning measure, which is based on Helmholtz principle from Gestalt Theory. In SMC, meaningfulness of terms in the context of classes are calculated and used for classification of the document.

Dengya Zhu et all (2011) [2] Term weighting strategy plays an important role in the areas related to text processing such as text categorization and information retrieval. In such systems, term frequency, inverse document frequency, and document length normalization are important factors which are to be considered when a term weighting strategy is developed. Term length normalization is proposed to give equal opportunities to retrieve both lengthy documen ts and shorter ones.

Nam Do et all (2012) [3] This link model is incorporated with the Markov Random Field model to form the soft labeling model for text classification. This new approach has combined both the local content and the influence from the neighborhood. The results of soft labeling model on standard data sets are also promising. Moreover, the new model can be applied not only on the text classification problem but also many kinds of richly structured data sets.

Haydemar Nunez et all (2012) [4] In this work an automatic classifier of undergraduate final projects based on text mining is presented. The dataset, comprising documents from four professional categories, was represented by means the vector space model with different index metrics. Also, a number of techniques for reduction dimensionality were applied over the word space. In order to construct the classification model the K-nearest neighbor algorithm was applied. Using 10-fold cross-validations we could obtain 82% of predictive accuracy.

Devendra Singh Rathore et all (2013) [5] Selection of research projects is an important research topic in research and development (R&D) project management. With the ever-increasing quantity of text data from a variety of online sources, it is a significant task to categorize or classify these text documents into categories that are manageable and easy to understand.

Seema Singh et all (2014) [6] Text categorization task have gained the attention of researchers in last 10 years with the increase in web-based contents of documents. For searching a particular document from the web or any large document collection text or document categorization is most useful task. We demand some better system and enhanced machine learning classifiers to accomplish task of document categorization. We designed a multi-agent based system that consists of some software hybrid agents that obtains the category of a document and interact with each other to take final decision about the category and then data is fed to a machine learn ing classifier in order to enhance the performance.

Dai Li et all (2014) [7] The HPHR system was evaluated on documents drawn from two different applications, vehicle fault diagnostic documents, which are in a form of unstructured and verbatim text descriptions, and Reuters corpus. The performance of the proposed system, HPHR, on both document collections showed superiority over the systems commonly used in text document.

R.Yasotha et all (2015) [8] There are two approaches, rule-based and machine learning-based, that are used to automate classification task. Motivated by its limitations, this paper proposes a Latent Dirichlet Allocation (LDA) based approach to automatically classify text documents. In order to develop and test the proposed approach on a realistic set up, ACM (Association for Computing Machinery) Computing Classification System (CCS) is selected as the target platform and 9100 computer science related articles categorized under ACMCCS were selected.

P.Anupriya et all (2015) [9] In this work, a dataset with 200 abstracts fall under four topics are collected from two different domain journals for tagging journal abstracts. T he document models are built using LDA (Latent Dirichlet Allocation) with Collapsed Variational Bayes and Gibbs sampling. Then the built model is used to extract appropriate tags for abstracts.

S.Subbaiah et all (2013) [10] It uses ODP taxonomy and domain ontology and datasets to cluster and identify the category of the given text document. The proposed work has three steps, namely, preprocessing, rule generation and probability calculation. At the stage of preprocessing the input document is split into paragraphs and statements. In

rule generation, the documents from the training set are read. In probability calculation, positive and negative weight factor is calculated.

Jian-Ping Mei et all (2014) [11] In this study, we work on clustering approaches that take care of both the large-scale and high-dimensionality issues. Specifically, we propose two methods for incrementally clustering of document data. The first method is a modification of the existing FCM-based incremental clustering with a step to normalize the centroids in each iteration, while the other method is incremental clustering.

Zhiyang He et all (2015) [12] Multilabel categorization has received a great deal of attention in recent years. This paper proposes a novel probabilistic generative model, label correlation mixture model (LCMM), to depict the multiply labeled documents, which can be used for multilabel spoken document categorization as well as multilabel text categorization. In LCMM, labels and topics have the one-to-one correspondence. The LCMM consists of two important components: 1) a label correlation model and 2) a multilabel conditioned document model. The label correlation model formulates the generating process of labels where the dependences between the labels are taken into account.

Vani K et all (2014) [13] The paper focuses on application of automatic text document categorization in plagiarism detection domain.This paper aims on the study and comparison of different methods of document categorization in external plagiarism detection. The primary focus is to explore the unsupervised document categorization/ clustering methods using different variations of K-means algorithm and compare it with the general N-gram based method and Vector Space Model based method.

Heba Ayeldeen et all (2013) [14] The fuzzy Euclidean distance clustering algorithm is well studied and used in information retrieval society for clustering documents. In this paper we proposed results for clustering theses documents based on Euclidean distance and cluster dependent keyword weighting. The proposed approach is based on the Fuzzy Euclidean distance clustering algorithm.

Hao Lin et all (2014) [15] In this paper, we evaluate energy cost of different classifiers and reduce the energy cost by parallelization, trying to find a classifier that performs the best on both aspects – effectiveness and efficiency.

VI. CONCLUSION

In this paper, we surveyed the latest literature review on document categorization. We gave brief discussion of document categorization and its techniques like NB, SVM, k-NN etc. In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is the non-parametric method used for classification and regression. In both cases, the input consist of k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In future work we cannot work not only on k-NN but we can work on its advanced feature i.e. weighted k-NN classifier

.

REFERENCES

[1] Murat Can Ganiz , Melike Tutkan , Selim Akyokus, “A Novel Classifier Based on Meaning for Text Classification” 2015.

[2] Dengya Zhu , Jitian XIAO, “R-tfidf, a Variety of tf-idf Term Weighting Strategy in Document Categorization” 2011 Seventh International Conference on Semantics, Knowledge and Grids.

[3] Nam Do-Hoang Le Thai-Son Tran Minh-Triet Tran,”Exploring neighborhood influence in text classification” 2012 Fourth International Conference on Knowledge and Systems Engineering.

[4] Haydemar Nuriez, Esmeralda Ramos,”Automatic classification of academic documents using text mining techniques” 2012 IEEE.

[5] Devendra Singh Rathore, Dr. R.C. Jain, Babita Ujjainiya,”A Text Mining Method for Research Project Selection using KNN” 2013 IEEE.

[6] Seema Singh, Chandra Prakash,”Document Categorization in Multi-Agent Environment with Enhanced Machine Learning Classifier” 2014 IEEE.

[7] Dai Li, Yi L. Murphey,”Automatic Text Categorization Using a System of High-Precision and High-Recall Models” 2014 IEEE.

[8] R. Yasotha, E. Y. A. Charle,” Automated Text Document Categorization” 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS'15).

[9] P. Anupriya, S. Karpagavalli,”LDA Based Topic Modeling of Journal Abstracts”2015 International Conference on Advanced Computing and Communication Systems (ICACCS -2015), Jan. 05 – 07, 2015, Coimbatore, India.

[10] S.Subbaiah,"Extracting Knowledge using Probabilistic Classifier for Text Mining" Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22
[11] Jian-Ping Mei, Yangtao Wang, Lihui Chen,Chunyan Miao,"Incremental fuzzy clustering for document categorization" 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) July 6-11, 2014, Beijing, China.
[12] Zhiyang He, Ji Wu, Tao Li,"Label Correlation Mixture Model: A Supervised Generative Approach to Multilabel Spoken Document Categorization"Digital Object Identifier 10.1109/TETC.2014.2377559
[13] Vani K, Deepa Gupta,"Using K-means Cluster Based Techniques in External Plagiarism Detection" 2014 IEEE
[14] Reba Ayeldeen,Aboul Ella Rassanien,Aly Aly Fahm,"Fuzzy clustering and categorization of text document" 2013 IEEE
[15] Hao Lin,"Research on energy-efficient text classification" 2nd International Conference on Information Technology and Electronic Commerce (ICITEC 2014) .