RESEARCH ARTICLE OPEN ACCESS

Machine Learning For Real Estate Contracts Automatic Categorization of Text

Mr.C.Mani M.C.A.,M.Phil.,M.E [1], J.Jayasudha[2]
Assistant professor[1], Research Scholar[2] ,
Department of Computer Applications,Nandha Engineering College/Anna University,
Erode.

--************************----------------------------------

Abstract:

Automatic Text Classification is a machine learning task that automatically assigns a given document to a set of pre-defined categories based on its textual content and mined features. Automatic Text Classification has important applications in content management, contextual search, estimation mining, product review analysis, spam filtering and text sentiment mining. This paper explains the generic strategy for automatic text classification and analyses existing solutions to major issues such as dealing with unstructured text, handling large number of features and selecting a machine learning technique appropriate to the text-classification application.

There are statistical model, rule based model, hybrid model. Statistical model is based on training text which configured in each categories, Rule Based model is based on rules like Positive term, Negative term, Relevant term, Irrelevant term. Positive term list of mandatory terms. Negative Term list of excluding terms. Relevant Term list of relevant terms. Irrelevant Term list of irrelevant terms. Hybrid model is combination of statistical and rule based model. Hybrid model will give the accurate result. At first model will be created as statistical model to get the exact result later for fine tuning process have to add terms so at last the model will look as hybrid model.

We will discuss in detail issues pertaining to three different problems, namely, document representation, classifier construction, and classifier evaluation.

Keywords — **Artificial Intelligence, Machine Learning, Mining, Automatic text classification, feature extraction, pre-processing, text mining, Natural Language Processing**

--************************--------------------------------

1. INTRODUCTION

Automatic Text Classification involves conveying a text document to a set of pre-defined classes automatically, using a machine learning technique. For example lease administration to classifying the classes like Alteration,Task and Subletting, Audit Rights, Default, Holdover, Repairand,Maintenance,Insurance,Parking,Utilities,Surrender and restoration, late fee. These kind of clauses to classified the lease document.

The organization is usually done on the basis of important words or structures removed from the text document. Then the classes are pre-defined it is a gifted machine learning task. Most of the official communication and documentation maintained in profitable and real estate organizations is in the form of word-based electronic documents and electronic mail. Considerable of the personal and other communication complete by private characters is in the form of e-mails, blogs etc. Owing to this information overload, efficient classification and retrieval of pertinent content has gained significant importance.

This paper explains the common approach for automatic text organization which includes steps such as pre-processing (eliminating stop-words [1] [2] [3], stemming [2] [4] etc.), eye collection using several statistical or semantic approaches, and modelling using suitable machine learning techniques.

About of the major subjects involved in automatic text organization such as commerce with unstructured text, action large number of attributes, investigative success of purely statistical pre-processing systems for text organization v/s semantic then natural language dispensation based techniques, dealing with absent metadata and choice of a apposite machine learning method for training a text classifier.

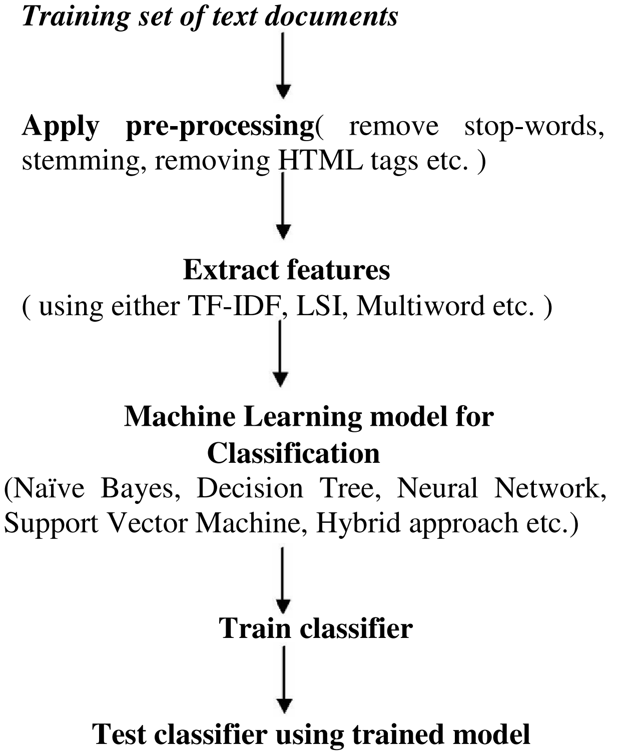
2. GENERIC STRATEGY FOR CLASSIFYING A TEXT DOCUMENT

The generic strategy for text classification is portrayed. The main steps involved are i) document pre-processing, ii) feature extraction / selection, iii) model selection, iv) training and testing the classifier.Data pre-processing decreases the size of the input text documents meaningfully. It involves activities like sentence boundary determination [2], natural language specific stop-word elimination [1] [2] [3] and stemming [2] [4].extracting text, tokenization, stop words removal and lemmatization required for automatic classification.

Stop-words are functional words which occur frequently in the language of the text (for example, „a‟, ‟the‟, ‟an‟, ‟of‟ etc. in English language), so that they are not useful for society. Stopping is the action of plummeting words to their root or base procedure.Aimed at English linguistic, the Porter‟s stemmer is a popular algorithm [4] [12], which is a suffix undressing sequence of methodical steps for stemming an English word, plunging the vocabulary of the training text by about one-third of its original size [4].

Feature removal / selection helps identify important arguments in a text document. This is done by methods like TF-IDF (term frequency-inverse document frequency) , LSI (latent semantic indexing), multi-word [2]etc. In the setting of text organization, features or attributes usually mean important words, multi-words or often occurring phrases indicative of the text category.

Afterward feature selection, the text document is signified as a document vector, then an suitable machine education algorithm is cast-off to train the text classifier. The qualified classifier is tested using a exam set of text documents. If the classification correctness of the trained classifier is found to be satisfactory for the test set, then this model is used to categorize new cases of text documents.


Training set of text documents
Apply pre-processing(remove stop-words, stemming, removing HTML tags etc.)
Extract features
(using either TF-IDF, LSI, Multiword etc.)
Machine Learning model for Classification
(Naïve Bayes, Decision Tree, Neural Network, Support Vector Machine, Hybrid approach etc.)
Train classifier
Test classifier using trained model

3. AUTOMATIC TEXT CLASSIFICATION

Instinctive text classification is a widely researched topic due to its applied applicability to numerous areas of text mining. The numerous issues in text organization and currently available solutions are discussed next. Retrieving metadata for classification and choice of mechanism learning method for text classification.

3.1.1. Classifying unstructured text

Some types of text documents like technical research papers are usually written severely in a pre-specified preparation, which makes it easier to classify them, because of positional info of attributes. Though, most text documents are written in an formless way, so classification has to be done on the basis of makings such as presence or absence of keywords and their incidence of incidence. Text documents can be signified as document courses using models such as the i)Multivariate Bernoulli Model [1]in which the document vector is a binary vector just representative the absence or attendance of feature footings; or the ii)Multinomial Model [1] in which document vectors additionally recollect the information concerning frequency of occurrence of nose terms.

3.1.2. Handling large number of attributes:Feature selection using statistical and semantic preprocessing techniques

Features useful in text classification are simple words from the language vocabulary, user-specified or extracted keywords, multi-words or metadata. In text organization literature, the steps complicated in feature discount are mostly applying pre-processing such as stop-word removal [1] [2] [3], stopping [4] etc. Text documents usually use words after a large vocabulary, nonetheless all words occurring in a document are not useful for classification. So, researchers consume proposed feature reduction techniques like TF-IDF,LSI[5],multi-word[2]etc. or a combination of such techniques. The TF-IDF is a chastely statistical method to evaluate the importance of a word based on its incidence of incidence in the document and in its relevant corpus. The LSI and multi-word methods are semantics concerned with techniques which also attempt to overcome the two basic glitches in classification „polysemy‟ (one word having many distinct meanings) then „synonymy‟ (different words having same meaning). The LSI technique essentially tries to use the semantics in a document construction using SVD (Singular Value Decomposition) matrix operations. A multi-word is a sequence of successive words having a semantic meaning (for example, “Information Technology”,“Delhi Public School”, ”Computer Engineering Department”, “State Bank of India”).

Multi-words are useful in classification as well as disambiguation. Several methods can be used to extract multi-words from text such as the incidence approach [2], mutual information approach etc.

3.1.3. Retrieving metadata useful for classification

Information about metadata is useful in classification. Metadata useful in classification are keywords, proper nouns such as names of persons / places, document title, name of document author [13] etc. Web documents optionally uphold metadata using the “META” tags which is very useful in organization.

Metadata such as keywords are often assumed by users through search. A method for retrieving topographies (spatial and contextual) and removing metadata using choice tree model has remained proposed in [13].

3.1.4. Modeling: Selection of appropriate machine learning technique for classification of text documents

Various oversaw machine learning methods have been proposed in literature for the automatic classification of text documents such as Unexperienced Bayes [1],Neural Networks, SVM (Support Vector Machine), Decision Tree and also by combining approaches [12] .

No single technique is found to be superior to all others for all types of cataloguing. The Naïve Bayesian classifier is based on the supposition of provisional independence among qualities. It gives a probabilistic organization of a text document provided there are a adequate number of exercise cases of each category. Then the Naïve Bayesian method is purely statistical its application is straightforward and knowledge time is less, however, its presentation is not good for categories defined with very few attributes/ features. SVM is originate to be very real for 2class organization problems (for example, text document belongs/ not belongs to a specific category; view is classified as positive/negative) then it is difficult to spread to multi-class organization. A class-incremental SVM organization method has been future. A Result Tree can be generated using procedures like ID3 [27] or C4.5 [13]. Unlike Naïve Bayesian organization, Decision Tree organization does not assume individuality among its features. In a Decision Tree picture the association between attributes is stowed as links. Result tree can be used as a text classifier once here are relatively fewer number of attributes to reflect, however it becomes problematic to manage for large number of attributes.

Investigators have reported improved classification accuracy by uniting machine learning methods. In [12], the presentation of Neural System based text organization was better by assigning the likelihoods derived from Naïve Bayesian method as initial weights. In, Naïve Bayesian technique was used as a preprocessor for dimensionality discount shadowed by the SVM method for text classification. There is a need to trial with more such hybrid methods in order to derive the extreme benefits after mechanism learning algorithms then to achieve better organization results.

4. THE MACHINE LEARNING APPROACH TO TEXT CATEGORIZATION

In the '80s, the most general approach (at least in operational settings) for the formation of automatic text classifiers contained in manually building, by means of knowledge engineering (KE) methods, an expert scheme capable of taking TC de-cisions. Such an expert scheme would typically contain of a set of manually clear logical rules, one per category, of type **if** h DNF formula *i* **then** h*category*i*:* A DNF ("disjunctive normal form") for-mula is a disjointedness of conjunctive clauses; the document is secret under h*category*i iff it satisfies the formulation, that is, iff it satisfies at least one of the clauses. The most well-known example of this method is the CONSTRUE system [Hayes et al. 1990], built by Carnegie Group for the Reuters news agency. A example rule of the type used in CONSTRUE is illustrated.

The disadvantage of this tactic is the knowledge acquisition bottleneck well recognized from the skilled systems literature. That is, the rubrics must be physically de-fined by a information engineer with the assistance of a area expert (in this case, an skilled in the association of documents in the chosen set of categories): if the set of groups is updated, before these two pro-fessionals must interfere again, and if the classifier is ported to a totally differ-ent area (i.e., set of categories), a differ-ent domain skilled needs to intervene and the work has to be recurrent from scratch. On the other hand, it was initially optional that this method can give very good efficiency results: Hayes et al. [1990] stated a .90 "breakeven" result (see Section 7) on a subset of the Reuters test collection, a number that outdoes classifier already exists and the original set of categories is updated, or if the clas-sifier is ported to a totally different area.

In the ML tactic, the reclassified documents are then the key reserve. In the most favourable case, they are al-ready obtainable; this classically happens for governments which have previously car-ried out the same classification activity physically and decide to mechanize the pro-cess. The less favorable case is once no physically classified leaflets are avail-able; this classically occurs for organi-zations that start a classification activ-ity and opt for an automatic modality immediately. The ML method is more suitable than the KE method also in this last case. In detail, it is easier to man-ually categorize a set of documents than to shape and tune a set of rubrics, meanwhile it is easier to describe a concept extension-ally (i.e., to select examples of it) than in-tensionally (i.e., to describe the concept in arguments, or to describe a process for rec-ognizing its instances).Classifiers built by income of ML tech-niques nowadays achieve impressive lev-els of effectiveness (see Section 7), making automatic organization a qualitatively (and not only economically) viable alter-native to physical classification.

The leaflets in *Te* cannot contribute in any way in the inductive building of the classifiers; if this complaint were not content, the new re-sults gotten would likely be unrealis-tically decent, then the valuation would thus have no technical character [Mitchell 1996, page 129]. In an employed setting, after assessment has been performed one would typically retrain the classifier on the entire initial corpus, in order to boost efficiency. In this case, the results of the previous assessment would be a pes-simistic approximation of the real presentation, since the final classifier has been skilled on more data than the classifier assessed.

This is called the train-and-test ap-proach. An another is the k-fold cross-validation method (see Mitchell [1996], page 146), in which *k* dissimilar classi-fiers 8_1, : : : , 8_k are constructed by partition-ing the initial corpus into *k* disjoint sets Te_1, : : : , Te_k and then iteratively apply-ing the train-and-test approach on pairs h$T\ V_i$ D $-Te_i$, Te_i i. The final effectiveness numeral is obtained by separately comput-ing the effectiveness of 8_1, : : : , 8_k , and then averaging the separate results in some way.

In both methods, it is often the case that the interior parameters of the clas-sifiers necessity be tuned by challenging which standards of the limits yield the best efficiency. In order to brand this op-timization possible, in the train-and-test method the set fd_1, : : : , d $_{j\ T\ V\ j}$g is furthersplit into a exercise *set Tr* D fd_1, : : : , d_{jTrj}g, after which the classifier is built, and a *val-idation set Va* D fd_{jTrjC1}, : : : , $d_{jT\ V\ j}$g (sometimes called a *hold-out set*), on which the recurrent tests of the classifier aimed even the finest classifiers built in the late ’90s by state of-the-art ML techniques. Though, no other classifier has been verified on the same dataset as CONSTRUE, and it is not clear whether this was a arbitrarily chosen or a favourable subset of the entire Reuters group. As argued by Yang [1999], the results above do not let us to state that these effectiveness results may be got in general.

Meanwhile the early ’90s, the ML approach to TC has gained popularity and has finally become the leading one, at least in the research community (see Mitchell [1996] for a complete intro-duction to ML). In this method, a general inductive process (also called the *learner*) mechanically builds a classifier for a cat-egory c_i by detecting the physiognomies of a set of leaflets manually classified under c_i or c^{-}_{i} by a area expert; after these physiognomies, the inductive procedure gleans the physiognomies that a new hidden text should have in order to be secret under c_i . In ML terminology, the organization problem is an action of *supervised* knowledge, meanwhile the knowledge procedure is “supervised” by the knowledge of the groups and of the exercise in-stances that fit to them.[2]

The compensations of the ML method over the KE method are evident. The en-gineering exertion goes to the construc-tion not of a classifier, nonetheless of an involuntary builder of classifiers (the *learner*). This income that if a beginner is (as it often is) available off-the-shelf, all that is needed is the inductive, *automatic* building of a classifier from a set of physically clas-sified leaflets. The same happens if a at limit optimization are did; the obvious irregular may be used in the *k*-fold cross-validation case. Note that, for the similar reason why we do not test a clas-sifier on the leaflets it has been skilled on, we do not test it on the documents it has been enhanced on: test set and vali-dation set necessity be kept separate.[3]

Assumed a corpus, one may describe the generality $g\ (c_i\ )$ of a category c_i as the fraction of leaflets that belong to c_i , that is: The exercise *set* generalization $g_{Tr}(c_i\ )$, valida-tion set generalization $g_{Va}(c_i\ )$, and *test set gen-erality* $g_{Te}(c_i\ )$ of c_i may be defined in the clear way.

5.CONCLUSIONS

In this paper we future the method of Text Categorization on web documents using text mining then information removal based on the classical summarization techniques. First web documents are pre-processed to found an organized data file, by recognizing feature terms like term frequency count and weight percentage of each term. New results shows, this approach of Text Categorization is more suitable for Informal English linguistic based web content where there is vast amount of data constructed in relaxed terms. This method has meaningfully reduced the query response time, improved the precision and degrees of relevancy. Upcoming work comprises the use of Meta information such as the structure of the document, patterns changes and evaluation for Text Categorization

ACKNOWLEDGMENT

This paper owes a lot to the proposals and con-structure censure of Norbert Fuhr and David Lewis. Thanks also to Umberto Straccia for commentaries on an earlier draft, to Evgeniy

Gabrilovich, Daniela Giorgetti, and Alessandro Moschitti for spotting mistakes in an previous draft, and to Alessandro Sperduti for many fruitful discussions.

REFERENCES

[1] Kim S., Han K., Rim H., and Myaeng S. H. 2006 . Some effective techniques for naïve bayes text classification. IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 14571466.

[2] Zhang W., Yoshida T., and Tang X. 2007. Text classification using multi-word features. In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524.

[3] Hao Lili., and Hao Lizhu. 2008. Automatic identification of stopwords in Chinese text classification. In proceedings of the IEEE international conference on Computer Science and Software Engineering, pp. 718 – 722.

[4] Porter M. F. 1980. An algorithm for suffix stripping. Program, 14 (3), pp. 130-137.

[5] Liu T., Chen Z., Zhang B., Ma W., and Wu G. 2004. Improving text classification using local latent semantic indexing. In proceedings of the 4th IEEE international conference on Data Mining, pp. 162-169.

[6] M. M. Saad Missen, and M.Boughanem. 2009. Using WordNet‟s semantic relations for opinion detection in blogs. ECIR 2009, LNCS 5478, pp. 729-733, Springer-Verlag Berlin Heidelberg.

[7] Balahur A., and Montoyo A.. 2008. A feature dependent method for opinion mining and classification. In proceedings of the IEEE international conference on Natural Language Processing and Knowledge Engineering, pp. 1-7.

[8] Zhao L., and Li C.. 2009. Ontology based opinion mining for movie reviews. KSEM 2009, LNAI 5914, pp. 204-214, SpringerVerlag Berlin Heidelberg.

[9] Durant K. T., Smith M. D. 2006. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection,. WebKDD 2006, LNAI 4811, pp. 187-206, Springer-Verlag Berlin Heidelberg.

[10] Polpinij J., and Ghose A. K. 2008. An ontology-based sentiment classification methodology for online consumer reviews. In proceedings of the IEEE international conference on Web Intelligence and Intelligent Agent Technology, pp. 518-524.

[11] Ng V., Dasgupta S., and S. M. Niaz Arifin. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In proceedings of the 21 st international conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics, pp. 611-618.

[12] Goyal R. D. 2007. Knowledge based neural network for text classification. In proceedings of the IEEE international conference on Granular Computing, pp. 542 – 547.

[13] Changuel S., Labroche N., and BouchonMeunier B. 2009. Automatic web pages author extraction. LNAI 5822, pp. 300311, Springer-Verlag Berlin Heidelberg.