Content Oriented Automatic Text Categorization

[1]N.Balasundaraganapathy, [2]P.Yuvarasu [3] Mr.Prabhakar
[1]*Faculty, M.C.A, Panimalar Engineering College, Chennai*
[2]*PG Scholar, M.C.A, Panimalar Engineering College, Chennai*
[1]*Faculty, M.C.A, Panimalar Engineering College, Chennai*

Abstract:

The project is to implement a web spam classifier, which given a web page, will analyze its features and try to determine whether the page is spam or not. The efficiency of the classifier will be compared to the results spam detection in the text datasets using Naïve Baye's classifier text representation is the task of transforming the content of a textual document into a vector in the term space so that the document could be recognized and classified by a computer or a classifier. Different terms (i.e. words, phrases, or any other indexing units used to identify the contents of a text) have different importance in a text. The term weighting methods assign appropriate weights to the terms to improve the performance of text categorization. In this study, the investigate several widely-used unsupervised (traditional) and supervised term weighting methods on benchmark data collections in combination with NLP and Clustering algorithms. In consideration of the distribution of relevant documents in the collection, the propose a new simple supervised term weighting method, i.e. tf.rf, to improve the terms' discriminating power for text categorization task. a consistently better performance while other supervised term weighting methods based on information theory or statistical metric perform the worst in all experiments. On the other hand, the popularly used tf.idf method has not shown a uniformly good performance in terms of different data sets.

I. Introduction

The aim of this paper is to propose deep parallelism may be established between and an\ Automatic Text Categorization (ATC) the transmission of information and its reliable recovery. The main objective of our research has been to investigate how and up to which extreme the document representation space can be compressed and what are the effects on final classification of this compression. The idea behind is to set a first step toward an optimal encoding of the category, carried by the document vectorial representation, Categories are predefined by some external mechanism (normally human) by establishing learning phase. This forms the machine learning paradigm (as opposed to the knowledge engineering approach) MNB is a probabilistic categorizer that assumes a document is a sequence of terms, each of them randomly chosen among the term vocabulary, independently from the rest of term events in the document. Besides its oversimplified Naïve Bayes basis, the function of transferring information (i.e., a message) from a source to a destination. the encoder/transmitter, the transmission channel, and the receiver/decoder. The encoder/ transmitter processes the source message into the encoded and transmitted messages. A classical digital communication system simplified model is represented In its raw form, a text document is a string of characters. Typically in ATC, a bag-of-words approach is adopted, which assumes that the document is an order-ignored sequence of words that can be represented vectorially. It is further assumed that the vocabulary used by a given document depends on the category or topic it belongs to.

II. Existing System

Some previous works using content and link based features to detect spam are mainly focused on quantitative features rather than qualitative analysis. Other works used automatic classifiers to detect link-based spam, checksums and word weighting techniques and proposed a real-time system for word spam classification by using HTTP response headers to extract several features.

III. Proposed System

In this technique, a several new proposals and qualitative features to improve word spam detection. They are based on a group of link-based features which checks reliability of links and a group of content based features extracted with the help of Language Model approach. Finally an automatic classifier is build and that combines both these of features, reaching a precision that improves the results of each type separately and those obtained by other proposals. Some of the considered features are related to the quality of the links in the page, behavior of standard search Engines, applied to the queries thus increasing the spam detection rate. applied the approaches to detect them by using the percentages to relate them by using two techniques join them and find the ratings and applied both the approach synchronized and detect the method by links to be used in all users that to be applied and detecting whether the word is spam or non spam

IV. Module description

1. **Text Categorization (Encoder)**
2. **Spam Log**
3. **Evaluation of Text Classifiers (Decoder)**

Text Categorization (Encoder)

The document is determined by a Category encoder, which is a random selector of words, modulated by the category (i.e., the selection of words is a random event different and characteristic of each category). For each category input in the encoder is characterized by a distinct alphabet the words used by the documents and the conditional probabilities of each element of this alphabet.

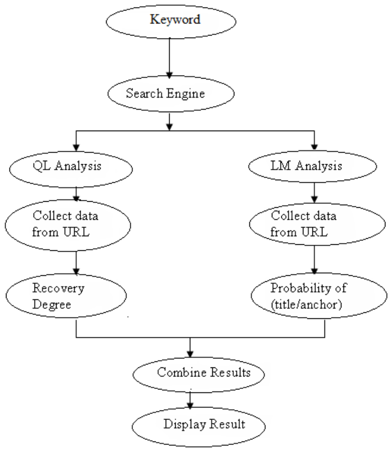
Spam Log

The results obtained from QLA Detection and LM Detection use Anchor Text (A), Surrounding anchor Text (S) and URL terms (U) as source of information. Purpose is to create two new sources of information: 1) combining Anchor Text and URL terms (AU) 2) combining Surrounding Anchor Text and URL terms. In addition, the user has considered other sources of information from the target page: Content Page (P), Title (T), and Meta Tags. From these various approaches user has applied to find out the divergence between this information. Various combination of the retrieved information helps out to find the spam more effectively. In many cases, user can find anchors with a small number of terms that sometimes mislead our results. However, by combining different sources of information such as Anchor text, Surrounding Anchor text, and URL terms, that can obtain a more descriptive language. Finally, user has combined content, link, LM, and QL features, achieving a more accurate classifier. All the Log details about the two analyses in the log for later verification and calculation of the spam.

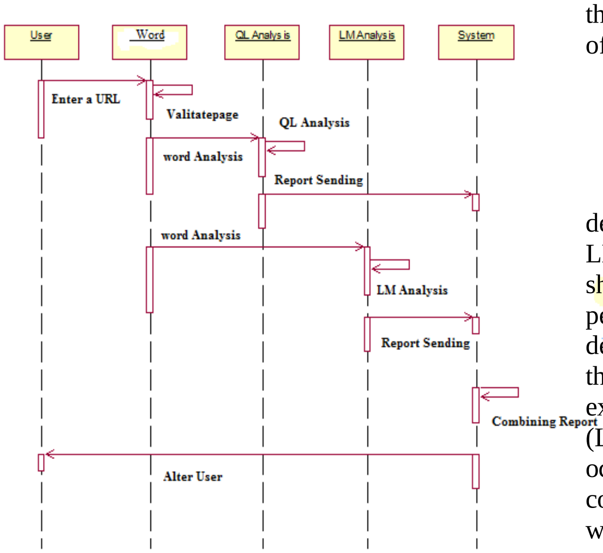
Evaluation Of Text Classifiers (Decoder)

The base for the token-category membership, spam filtering is now based on calculating the fuzzy similarity measure between the received message and each category, i.e. spam or legitimate. The message is then classified by comparing its fuzzy similarity measures. In order to calculate fuzzy similarity, user must first determine the membership degree of each token to the message. They are needed to determine the frequency of each token in the message and next membership degree is to be defined. The token with the maximum number of occurrences will be assigned a value of and all other tokens will be assigned proportional values

System Flow Diagram



Sequence Diagram



V. Method used

NLP

The Naive Bayes algorithm is a classification algorithm provided by Microsoft SQL Server Analysis Services for use in predictive modeling. The name Naive Bayes derives from the fact that the algorithm uses Bayes theorem but does not take into account dependencies that may exist, and therefore its assumptions are said to be naive.

This algorithm is less computationally intense than other Microsoft algorithms, and therefore is useful for quickly generating mining models to discover relationships between input columns and predictable columns. They are using this algorithm to do initial explorations of data, and then later you can apply the results to create additional mining models with other algorithms that are more computationally intense and more accurate.

A naive Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a naive Bayes classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature. Even though these features depend on the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability.

In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers

VI. Conclusion

In this project, A new methodology is proposed to detect spam in the Word, based on an analysis of QLs and LM. Therefore, comparisons with precompiled features show that the proposed methodology yields much better performance, indicating that LMs and QLs can be used to detect word spam effectively and in this (QLA) technique the approached is to find the recovery degree, using extraction of external and internal links. In language model (LM) method is to find the title occurrence & keyword occurrence with the use of anchor text. This also used to connect insert the word and with that use to connect the word file and type the original URLs site and in the spam process that are used to view the log details. These processes mainly used to find the word process that used in the search engine and the URLs typed to check whether the spam or non spam by comparing the percentages related to the appropriate limits used in the method. Thus the technique is use to find out the spam. This method proves to be more efficient in detecting the spam, word and non spam, word when compared to the previous models available for detecting spam.

VII. References

[1] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proc. 10th European Conf.Machine Learning (ECML), pp. 137-142, 1998.

[2] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher, “Spamrank— Fully automatic link spam detection,” in Proc. First Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb, Chiba, Japan, 2005, pp. 25–38.

[3] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna, “A reference collection for web spam,” *SIGIR Forum*,vol. 40, no. 2, pp. 11–24, 2006.

[4] T. Joachims, Learning to Classify Text Using Support Vector Machines—Methods, Theory, and Algorithms. Kluwer/Springer, 2002.

[5] T. M. Cover and J. A. Thomas, Elements of Information Theory. New York: Wiley-Interscience, 1991

[6] K. Bharat and M. R. Henzinger, “Improved algorithms for topic distillation in a hyperlinked environment,” in Proc. 21st Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval, New York, 1998, pp. 104–111, ACM.