Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 10, October 2015, pg.121 – 126

RESEARCH ARTICLE

A Study on Information Retrieval and Extraction for Text Data Words using Data Mining Classifier

S.C.Gowri[1], Dr. K.Meenakshi Sundaram[2]

[1]M.Phil Research Scholar, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, Tamilnadu, India

[2]Associate Professor, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, Tamilnadu, India

[1] gowriscmsc@gmail.com; [2] lecturerkms@yahoo.com

Abstract— Text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. In the current scenario, text classification gains lot of significance in processing and retrieval of text. Automated document classification becomes a key technology to deal and organize huge volume of documents and its frees organizations from the need of manually organizing document bases. A traditional approach to text categorization requires encoding documents into numerical vectors. This type of traditional document encoding causes two main problems: huge dimensionality and sparse distribution. Information retrieval techniques such as text indexing have been developed to handle the unstructured documents. The related task information extraction (IE) is about specific items in natural language documents.

Keywords— Information Retrieval (IR), Information Extraction (IE), Text Mining, Text Categorization

I. INTRODUCTION

The information has a great value and the amount of information has been expensive growing during last year's. Especially, text databases are rapidly growing due to the increasing amount of information available in electronic forms, such as electronic publications, e-mail, and the World Wide Web. Most of information in the world text hence the text categorization comes to the scene. Information retrieval (IR) is finding documents of an unstructured text that satisfies information need from within large collections (usually stored on computers). Information retrieval is fast becoming the dominant form of information access. Due to the rapid growth of text information, information retrieval system has found many applications such as on-line library systems, on-line document management systems, and the more recently developed Web search engines. Text mining techniques are used to extract this valuable information from the raw text data and then integrate to build a structured database. Information extraction (IE) is the task of automatically extracting structured information from unstructured or semi-structured text.

II. LITERATURE SURVEY

Barker et al. [1] proposed an algorithm to choose noun phrases from a document as phrases. The frequency of noun are the features used in this work. Noun phrases are extracted from a text using a base noun phrase skimmer and an off the-shelf online dictionary.

HaCohen-Kerner et al. [2] proposed a model for phrase extraction based on supervised machine learning and combinations of the baseline methods. They applied J48, an improved variant of C4.5 decision tree for feature combination.

Hulth et al. [3] proposed a phrase extraction algorithm in which a hierarchically organized thesaurus and the frequency analysis were integrated. The inductive logic programming has been used to combine evidences from frequency analysis and thesaurus.

A graph based model for key phrase extraction has been proposed by Y. Matsuo et al. [4]. A document is represented as a graph in which the nodes represent terms, and the edges represent the co-occurrence of terms. Whether a term is a keyword is determined by measuring its contribution to the graph.

J. Wang et al. [5] proposed a Neural Network based approach to phrase extraction that exploits traditional term frequency, inverted document frequency and position (binary) features. The neural network has been trained to classify a candidate phrase as key-phrase or not.

A key-phrase extraction program called Kea, developed by Frank et al. [6], uses the Bayesian learning technique for key-phrase extraction task. A model is learned from the training documents with exemplar key-phrases and corresponds to a specific corpus containing the training documents.

Kamal Sarkar et al. [7] proposed a neural network based model for key-phrase extraction by extracting noun phrases from a document. And the neural network has been trained to classify a candidate phrase as key-phrase or not.

Sebastiani [8] surveyed ten-years of previous research on text categorization. In this research the text categorization is a very important task especially for information retrieval, and recommends machine learning based approaches, rather than rule based approaches. The method mentioned the application of text categorization to document organization briefly, presented more than ten machine learning based approaches, and stated that there are more approaches in addition to those. His survey is very significant for this research as a tutorial on text categorization.

K Nearest Neighbor was initially created by Cover and Hart in 1967 as a genetic classification algorithm. It was initially applied to text categorization by Massand et al. at [9]. The K Nearest Neighbor algorithm is quite simple: given a test documents, and uses the categories of the K neighbors to weight the category candidates. K Nearest Neighbor is a lazy supervised learning algorithm, because it learns sample data selectively after unseen data is given for their classification. Unlike other supervised learning algorithm, it does not learn all the sample data in advance.

Naive Bayes may be considered as another approach to text categorization. It was initially created by Kononenko in 1989, based on Bayes Rule. Its application to text categorization was mentioned in the textbook by Mitchell. Assuming that the Naive Bayes is the popular approach, Mladenic et al. [10] proposed and evaluated feature selection methods.

XindongWu et al. [11] presented the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM): C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community With each algorithm, the provide a description of the algorithm, discuss the impact of the algorithm, and review current and further research on the algorithm.

Ada et al. [12] has presented the some data mining classification techniques such as neural network & SVMs for detection and classification of Lung Cancer in X-ray chest films. Due to high number of false positives extracted, a set of 160 features was calculated and a feature extraction technique was applied to select the best feature. They classify the digital X-ray films in two categories: normal and abnormal. The normal or negative ones are those characterizing a healthy patient. Abnormal or positive ones include types of lung cancer.

Anjali Ganesh Jivani [13] discussed that the purpose of stemming is to reduce different grammatical forms or word forms of a word like its noun, adjective, verb, adverb etc. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. This paper discusses different methods of stemming and their comparisons in terms of usage, advantages as well as limitations. The basic difference between stemming and lemmatization is also discussed.

Vishal Gupta et.al [14] has analyzed the stemmer‟s performance and effectiveness in applications such as spelling checker varies across languages. A typical simple stemmer algorithm involves removing suffixes using a list of frequent suffixes, while a more complex one would use morphological knowledge to derive a stem from the words. The paper gives a detailed outline of common stemming techniques and existing stemmers for Indian languages.

Sunita Beniwal et al. [15] used various methods for classification exists like bayesian, decision trees, rule based neural networks etc. Before applying any mining technique, irrelevant attributes needs to be filtered.

Filtering is done using different feature selection techniques like wrapper, filter, embedded technique. Also provide a survey of various feature selection techniques and classification techniques used for mining.

Li et al. [16] have used VIPS to segment Web pages. They form a feature-vector considering visual, spatial, and content aspects of the extracted blocks to cluster them later.

III.TEXT MINING

Text mining has been defined as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources". Text mining, which is sometimes referred to "text analytics", is one way to make qualitative or "unstructured" data usable by a computer. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification they are supervised, unsupervised, semi supervised and summarization. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases or XML files, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files.

3.1 Text mining Techniques

Text mining involves the application of techniques from areas such as information retrieval, natural language processing, and information extraction.

3.1.1 Information Retrieval

Information Retrieval (IR) systems identify the documents in a collection which match a user's query. Information retrieval is the task of obtaining relevant information from a collection of resources. It is used to focus on the textual information which includes text as well as document retrieval.

Document retrieval is measured as an extension of the information retrieval where the documents that are returned are processed to condense or extract the particular information sought by the user. To studies the retrieval of information from a collection of written text documents is called Information retrieval (IR).

It can reduce information overload by using automated information retrieval systems. The information retrieval mainly deals with the large range of information processing from information retrieval to knowledge retrieval. Information retrieval system is used in online digital library, online service and online document system and web search engines. There are other powerful techniques in text mining like classification, categorization and summarization, clustering to handle large amount of text data.

3.1.2 Information Extraction

Information Extraction (IE) is the Process of automatically obtaining structure data from an unstructured national language Document retrieval is measured as an extension of the information retrieval where the documents that are returned are processed to condense process of automatically obtaining structured data from an unstructured natural language document. Information extraction (IE) is the task of automatically extracting structured specific information from unstructured or semi-structured natural language text. Often this involves defining the general form of the information that interested in as one or more templates, which are then used to guide the extraction process. The main goal of information extraction is making information more accessible to the people and more machine-process able. There are two main problems Associates with IE.

- Paraphrase
- Paraphrase- many ways to say the same thing.
- Ambiguity-the same word/ phase/ sentence may mean different things in different contents.

Data integration which include the representation of an entity their relationship and large scale entity and relation resolution Natural Language Processing (NLP) is one of the oldest and most difficult problems in the field of artificial intelligence. It is the analysis of human language so that computers can understand natural languages as humans do. Although this goal is still some way off, NLP can perform some types of analysis with a high degree of success.

3.1.3 Text Categorization

Text categorization is one of the well studied problems in data mining and information retrieval. Categorization is the process in which ideas and objects are recognized, differentiated and understood. Categorization implies that objects are grouped into categories, usually for some specific purpose. A category illuminates a relationship between the subjects and objects of knowledge. The data categorization includes the categorization of text, image, object, voice etc. Text categorization becomes a key technology to deal with and organize large numbers of documents. Text categorization is the assignment of natural language documents to one or more predefined categories based on their semantic content is an important component in many information organization and management tasks. Automatic text categorization is treated as a supervised learning task. The goal of this task is to determine whether a given document belongs to the given category or not by looking at the synonyms or prefix of that category.

Fig.1.1 Process for encoding document

Text documents are classified based on rules, like expert systems, while in the latter type of approaches they are categorized based on classifiers trained with predefined documents. Machine learning based approaches have tended to replace rule based approaches, because machine learning based approaches are more flexible.

Decision Trees: Decision trees are designed with the use of a hierarchical division of the underlying data space with the use of different text features.

Neural Network Classifiers: Neural networks are used in a wide variety of domains for the purposes of classification. In the context of text data, the main difference for neural network classifiers is to adapt these classifiers with the use of word features. Note that neural network classifiers are related to SVM classifiers; indeed, they both are in the category of discriminative classifiers, which are in contrast with the generative classifiers.

Support Vector Machine: The first approach to classifier training and document classification is SVM (Support Vector Machine). This approach is based on a kernel based supervised learning algorithm for a binary classification problem. SVM defines two hyper planes corresponding to the positive and negative classes in the linear separable feature space mapped from the given original space using a kernel function with maximal margin between them. Support vectors correspond to the training samples that influence the determination of the two hyper planes as the boundaries of the two classes.

3.1.4 Applications of Text Classification

The advances from Information Retrieval and Artificial Intelligence have made document classification a hot issue. Document classification may appear in many applications:

Email Filtering: Systems for filtering a person's incoming Emails to weed out scam, or to categorize them into different classes, are just now becoming available (for example the Automatic Organizer by Intel).

Document Organization and Retrieval: The above application is generally useful for many applications beyond news filtering and organization. A variety of supervised methods may be used for document organization in many domains.

Opinion Mining: Customer reviews or opinions are often short text documents which can be mined to determine useful information from the review.

Enterprise Business Intelligence: Enterprise business intelligence is the deployment of BI throughout an enterprise, usually through the combination of an enterprise data warehouse and an enterprise license to a BI platform or tool set that can be used by business users in various roles.

Security applications: Many text mining software packages are marketed for security applications, especially monitoring and analysis of online plain text sources such as Internet news, blogs, etc. for national security purposes. It is also involved in the study of text encryption/decryption.

Biomedical applications: One online text mining application in the biomedical literature is GoPubMed. GoPubmed was the first semantic search engine on the Web. Another example is PubGene that combines biomedical text mining with network visualization as an Internet service.

Online media applications: Text mining is being used by large media companies, such as the Tribune Company, to clarify information and to provide readers with greater search experiences, which in turn increases site "stickiness" and revenue.

Sentiment analysis: Sentiment analysis may involve analysis of movie reviews for estimating how favourable a review is for a movie. Such an analysis may need a labelled data set or labelling of the affectivity of words.

IV. EXPERIMENTATION & RESULTS

Microsoft included a new language called C# (pronounced C Sharp). C# is designed to be a simple, modern, general-purpose, object-oriented programming language, borrowing key concepts from several other languages, most notably Java. C# could theoretically be compiled to machine code, but in real life, it's always used in combination with the .NET framework. Therefore, applications written in C#, requires the .NET framework to be installed on the computer running the application.

Precision and recall are widely used for evaluation measures in text categorization. Precision is a measure of the accuracy provided that a specific class has been predicted. Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. An information retrieval system often

needs to trade off recall for precision or vice versa. One commonly used tradeoff is the F-score, which is defined as the harmonic mean of recall and precision.

Techniques	Precision	Recall
Decision Trees	0.48	0.52
Neural Network Classifiers	0.33	0.67
Association Rule Mining	0.63	0.37

TABLE 1 .1 RESULTS OF PRECISION AND RECALL

The above table 1.1 shows the complete recall and precision rate on given document collection. In SVM experiment, precision and recall are low in some categories it takes more training time than neural based classifier model. From the above shown the experimental results of the proposed methodology and the existing system among the obtained results the proposed methodology is pictorial better recognition rate by sampling with the existing methodology using neural network. The pictorial representation of the proposed methods by comparing with the existing methodology presented below.

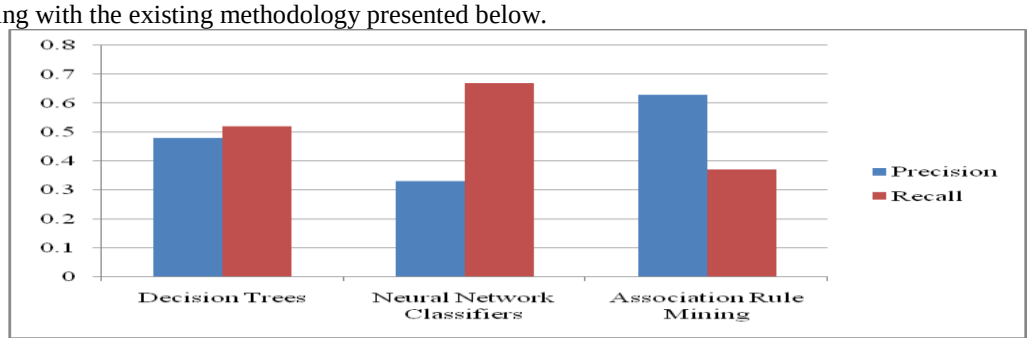

Fig1.2 PERFORMANCE EVALUATION

From the above figure 1.2 shows the graphical representation of the result which is different in the table 1.1 result of proposed method and existing method of valuable feature extraction for efficient classification to obtain reducing training and testing iteration in terms of document classification. The result shows the performance of proposed method in terms Precision and recall values for information retrieval of documents.

V. CONCLUSION

In this paper discussed the various methods for document representation used a full inverted index as the basis for the operation on string vectors. It is better than performing traditional approach to represent the document by minimizing document pre-processing time and feature dimensionality also it provides the potential easiness for tracing why each document is classified under the category. In this research work to suggest the encode documents into string vectors by extracting phrase than into numerical vectors.

ACKNOWLEDGEMENT

My heartfelt gratitude goes to my beloved guide Dr.K.Meenakshisundaram Associate Professor of Computer Science, Erode Arts and Science College (Autonomous), Erode for his valuable guidance support my research work successfully. I express my sincere thanks to HOD, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode for the encouragement to complete the study in a successful manner. My warm acknowledgement to all Faculty Members of my department for their inspiring support and constant encouragement during the research activities.

REFERENCES

[1] K. Barker, N. Cornacchia, Using Noun Phrase Heads to Extract Document Keyphrases. In H. Hamilton, Q. Yang (eds.): Canadian AI 2000. Lecture Notes in Artificial Intelligence, 2000, Vol. 1822, Springer-Verlag, Berlin Heidelberg, 40 – 52.

[2] Y. HaCohen-Kerner, Automatic Extraction of Keywords from Abstracts, In V. Palade, R. J. Howlett, L. C. Jain (eds.): KES 2003. Lecture Notes in Artificial Intelligence, 2003, Vol. 2773,Springer-Verlag, Berlin Heidelberg, 843 – 849.

[3] Y. Matsuo, Y. Ohsawa, M. Ishizuka, KeyWorld: Extracting Keywords from a Document as a Small World, In K. P. Jantke, A. shinohara (eds.): DS 2001. Lecture Notes in Computer Science, 2001, Vol. 2226, Springer-Verlag, Berlin Heidelberg, 271– 281.

[4] J. Wang, H. Peng, J.-S. Hu, Automatic Keyphrases Extraction from Document Using Neural Network., ICMLC 2005, 633-641.

[5] E. Frank, I. H. Witten, G.W. Paynter, KEA: Practical Automatic Keyphrase Extraction, In E. A. Fox, N. Rowe (eds.): Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries. 1999, ACM Press, Berkeley, CA , 254 – 255.
[6] Kamal Sarkar, Mita Nasipuri and Suranjan Ghose "A New Approach to Keyphrase Extraction Using Neural Networks" International Journal of Computer Science Issues, Vol. 7, Issue 2, No 3, March 2010.
[7] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Survey, Vol.34, No.1, pp.1-47, 2002.
[8] Massand, B., Linoff, G., Waltz, D. (1992). Classifying News Stories using Memory based Reasoning, In: the Proceedings of 15th ACM International Conference on Research and Development in Information Retrieval, p. 59-65.
[9] Mladenic, D., Grobelink, M. (1999). Feature Selection for unbalanced class distribution and Naïve Bayes. In: the Proceedings of International Conference on Machine Learning, p. 256-267.
[10] XindongWu, Vipin Kumar, J.Ross Quinlan, Joydeep Ghosh, Qiang Yang , Top 10 algorithms in data mining, Knowl Inf Syst (2008) 14:1–37.
[11] Ada, Rajneet Kaur. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 3, March 2013 ISSN: 2277 128X.
[12] Anjali Ganesh Jivani , A Comparative Study of Stemming Algorithms, International Journal of Computer, Technology and Application, Volume 2, ISSN:2229-6093,2010.
[13] Vishal Gupta, Gurpreet Singh Lehal, A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages, Journal of Emerging Techniloigies in Web Intelligence, VOL. 5, NO. 2, MAY 2013.
[14] Sunita Beniwal, Jitender Arora. Classification and Feature Selection Techniques in Data Mining.International Journal of Engineering Research & Technology (IJERT). Vol. 1 Issue 6, August - 2012. ISSN:2278-0181.
[15] C. Li, J. Dong, J. Chen. Extraction of informative blocks from Web pages based on VIPS. Journal of Computational Information Systems 6 (1) ,271–277,2010.

