ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

A STUDY ON RELEVANT FEATURE EXTRACTION ON TEXT MINING USING DM CLASSIFICATION TECHNIQUES

C.Malarvizhi[#1], R. Nivedhitha[#2]

#C.Malarvizhi[1]
Research Scholar,
Department of Computer Science,
Avinashilingam University,
Coimbatore, India
malarchan03@gmail.com

#R. Nivedhitha[2]
Research Scholar,
Department of Computer Science,
Avinashilingam University,
Coimbatore, India
nivedhirengaraj@gmail.com

ABSTRACT: Fastest upgrading technology of Text mining, text analysis is a most viewable one by every researcher. Today a text analysis is the important field for many areas like hospital, education; web search etc., extracting the text for analyzing the relevant value is an important factor in today's world. By extracting the text is an interesting knowledge of which is relevant or irrelevant for those fields in various text documents. Nowadays it is the big challenging for those fields for extracting the text. However many researcher were found many techniques for solving this issues. Some of the technique was succeed and some provide drawbacks. This paper shows the various classification techniques in data mining to extract the relevant feature in the text documents.

Keywords: Data mining, Text Mining, Text document, Relevant Feature extraction, Classification technique

I Introduction

With the speedy growth of text communication available in the internet, user wants to share the information in both formatted and unformatted manner. In the method of formatted text, it has the extra features like text style, size, color, etc. but in the form of unformatted text it is like a plain text. To extract the relevant feature of text is called a text mining mechanism, Text mining is the finding of attractive familiarity in text documents. It is a difficult issue to discover wonderful knowledge (or features) in text documents. Text mining is an information finding method that manages to pay for computational intelligence. Whereas mining the text the user have to scrutinize the text content in the text document. The text analysis is one of the vital factors in this real world to analyze the associated features. Text Analytics articulate a set of languages, mathematics, and machine learning methods that shape and form the information fulfilled of textual cause for business intelligence [1]. Text mining is the discovery of extraordinary knowledge in text documents. It's complicated matter to find out appropriate data in text documents to help users to search for out what they want. It is a challenging work to use those outlines and also fetch them up to date. Earlier expression based methods are make available by Information Retrieval (IR) techniques. All expression based process endures from troubles such as polysemy and synonymy. When a word has a multiplicity of meanings, it is known as polysemy. When different words have the correspondent meaning, it is called synonymy [2]. Thus the semantic meaning of an assortment of discovered terms are changeable for respond what users crave. The rationale of text mining is to process shapeless data, mine meaningful information in text collection.

Information can be extorts to derive summaries for the words restricted in the documents. Hence, we can investigate documents and determine comparison between them. Text mining also referred as text data mining, approximately corresponding to text analytics, it refers to practice of deriving high superiority of information from text and high quality of information is resultant from end to end devising of patterns. Text analysis engrosses information retrieval, lexical analysis, word occurrence distributions, pattern acknowledgment, information extraction and relevant feature extraction [3]. The significance feature discovery (RFD) is to locate the constructive features accessible in text documents, including both relevant and unrelated ones, for recitation text mining results. This is a predominantly challenging task in contemporary information scrutiny, from both an experimental and theoretical perspective. There were many issues in using pattern mining techniques for finding significance features in both relevant and irrelevant documents. The harms are the low-support problem. Given a subject, long patterns are usually more detailed for the topic, but they typically appear in documents with low maintain or frequency. If the smallest amount support is diminished a lot of raucous patterns can be exposed. And also in the misunderstanding problem, which means the measures (e.g., “support” and “confidence”) used in pattern mining turn out to be not appropriate in using patterns for answering this problems. For example, a highly recurrent pattern (in general a short pattern) may be a general pattern because it can be normally used in both related and unrelated documents. Hence, the tricky problem is how to use exposed patterns to truthfully weight practical features. Relevant Feature pulling out is used to mine the associated content in the database. The key goal of relevant feature detection is to discover useful features obtainable in a training set, counting both positive and negative documents, for recitation what users want. To finding the significance feature it have three approaches first, it is to adjust feature terms which have come into sight in both positive and negative example Secondly how often it will become visible in positive and negative credentials and finally portray which feature has the positive outline [4, 5]. For this examination this paper affords the study of different feature selection technique using classification in data mining.

II LITERATURE REVIEW

Yuefeng Li et al [6] analyze the existing popular text mining and categorization processes have assumed term based approaches. Nevertheless, they have all experience from the problems of polysemy and synonymy. Over the years, people have frequently held the suggestion that pattern-based process should execute superior than term-based ones in unfolding user partiality but many experimentation do not sustain this hypothesis. The pioneering technique obtainable in paper makes penetrate for this complicatedness. This method determine both positive and negative patterns in text documents as advanced level features in classify to truthfully weight low-level features (terms) based on their specificity and their distributions in the superior level features. Extensive experimentation using this technique on Reuters Corpus Volume 1 and TREC topics show that the proposed approach drastically outperforms both the state-of-the-art term-based methods underpinned.

Ning Zhong et al [7] examine many data mining techniques have been projected for mining useful prototype in text documents. Nevertheless, how to successfully use and inform exposed prototype is still an open research issue, particularly in the domain of text mining. While most existing text mining technique assumed term-based approaches, they all experience from the problems of polysemy and synonymy. Over the years, people have frequently held the hypothesis that outline (or phrase)-based approach should execute superior than the term-based ones, but many experimentation do not hold this hypothesis. This paper present an inventive and successful pattern sighting technique which includes the processes of example deploying and pattern growing, to progress the success of using and modernize discovered patterns for verdict relevant and motivating information. Significant experiments on RCV1 data anthology and TREC topics make obvious that the projected explanation achieves encouraging performance.

C.Kanakalakshmi and Dr.R.Manicka chezian [8] investigate the pulling out of functional information from shapeless textual data through the recognition and searching of motivating patterns. The detection of applicable features in real-world data for relating user information requirements or predilection is a new confront in text mining. Relevance of a feature designate that the features is at all times necessary for an most favorable subset, it cannot be unconcerned without upsetting the innovative provisional class sharing. They proposed an adaptive method for relevance feature detection is conversed to find useful features obtainable in a feedback set, as well as both positive and negative documents, for recitation what users need. Thus, this paper talk about the methods for relevance feature discovery using the replicated annealing rough calculation and genetic algorithm, a population of applicant solutions to an optimization problem on the way to better solutions. Harpreet Kaur, Rupinder Kaur et al [9] observed that the text mining using the pattern discovery usually uses only the text substance in standard fonts i.e. it does not believe the bold, underline or italic or even the larger fonts as the key text prototype for text mining. This generates difficulty many a times when the key words are remove from the article by the algorithm itself. In that case, significant keywords are left from the most important stream of text patterns. In their projected work, patterns are excavates in both positive and negative feedback. It then mechanically classifies the patterns into clusters to find pertinent patterns as well as eradicate noisy patterns for a given topic. A novel prototype organizing approach is proposed to remove choice features of text documents and use them for humanizing the retrieval performance. The projected approach is appraised by remove features from RF to progress the presentation of information filtering (IF).

Muthuvalli.A.R and Manikandan.M [10] study the feature clustering method to mechanically group terms into the three group positive specific features, general features, and negative specific features. The first issue in using irrelevant documents is how to choose a appropriate set of irrelevant documents since a very large set of negative example is characteristically get hold of For example, a Google Search can return millions of documents; though, only a few of those documents may be of attention to a Web user. Perceptibly, it is not well-organized to use all of the irrelevant documents. This representation is a supervised advance that needs a preparation set including both relevant documents and irrelevant documents. It also makes available recommendations for wrongdoer (irrelevant) collection and the use of specific stipulations and all-purpose terms for recitation user information needs. This model finds both positive and negative advice and the RFD used immaterial documents in the preparation set in order to eliminate the noises and also it can achieve the reasonable performance.

III FEATURE EXTRACTION USING CLASSIFICATION TECHNIQUE

Feature selection is the progression of selecting a subset of feature used to signify the data. In text categorization it spotlight on recognizing relevant information lacking touching the accurateness of the classifier. In text documents feature can be term, pattern, and condemnation. However, the traditional feature assortment methods are not successful for choosing text features for answering the relevance subject because significance is a solitary class problem [11, 12]. Investigate and solving the problem this study paper afford a various classification technique to overcome the issues. The classification technique we used in this paper is

a. K-Nearest Neighbor

KNN is a case based erudition algorithm. The objects are classified by choosing numerous labeled terms with their minimum distance from each object. The Major difficulty of KNN is that it uses all features in calculating reserve and costs very much time for categorizing objects. The classification is frequently executing by contrast the class frequencies of the k adjoining documents [13]. The assessment is done by calculating of angle among the two feature vectors. Feature vectors have to be regularizing to length 1. The main benefit of the k-nearest neighbor method is its straightforwardness. Its drawback is that it requires more time for classifying substance when there is a large amount of training examples. It is used to extract the nearest text in the text document and it analyzes the significant text data using the given set of attributes in the field.

b. Decision Trees Decision tree methods will renovate the manual classification of the documents by create well-defined queries (true/false) in the form of a tree structure where the nodes symbolize the questions and the leaves characterize their equivalent category of the documents. After the tree is twisted, a new document can be effortlessly be classified by situate them in to root node of the tree and run from side to side their query organization awaiting certain leaf is reached. The benefit of decision trees is that the production tree is easy to recognize even for persons who are not recognizable with the particulars of the model. The organization of tree is produce by the representation which makes available the user with combined view of the classification logic. A danger of the application of tree technique is "over fitting" that is if a tree more than fits then training data will classifies the training data bad but it would classify the documents to be categorized later better.

c. Support Vector-based Methods

There are two types of vector-based methods Centroid algorithm and Support vector machines. One of the simplest technique is the centroid algorithm. Throughout the learning stage standard feature vector for each group is calculated and it will be set as centroid-vector for the each sort. A document is with no trouble classify by determine the centroid-vector closest to its feature vector. The method is also shocking when the number of grouping is very large. Support vector machines (SVM) need positive training documents and also a positive number of negative preparation documents. SVM is looking for the conclusion surface that disconnect the positive term from the unhelpful examples in the n-dimensional space. The document is contiguous to the decision surface are called support vectors. The algorithm consequences remain unaffected if documents that not belongs to the support vectors and they are uninvolved from the training dataset. An advantage of SVM is runtime-behavior during the categorization of new documents. A drawback is that a document is assigned to various categories because of the resemblance calculated individually for each category.

IV COMPARISON ANALYSES

TECHNIQUES	**ADVANTAGES**	**DISADVANTAGES**
K-Nearest Neighbor	It is simple and effective Extract with the help of distance	Failed in long distances
Decision Trees	It assigns the accurate feature, It extracts in two bases which is positive and which is negative.	Classifies depend upon the text value
Support Vector-based Methods	It separates positive and negative data. It removed the noise data	It calculated individually

Table 1: Comparison analyses using classification techniques

V CONCLUSION

World provides a lot of information for the information retrieval, while retrieving the data most of the content were related to each other. They were provided both relevant and irrelevant data. Nowadays, it is the biggest challenge to discover the relevance feature in text document which helps to decide whether document is relevant or irrelevant. By extracting the text from them is a major issue, for these issues many researchers provide various techniques, this paper we have presented the study of relevant feature extraction using classification techniques in data mining. Many techniques were providing various functionalities; here we used K-Nearest Neighbor, Decision Trees and Support Vector Machine. From our analyses SVM provide a better classification function for text analyses of relevant extraction.

VI REFERENCES

[1] Priyanka R. Magar, C. S. Biradar, “Discovering Efficient Patterns for Text Mining Approach: A Survey”

[2] X. Li and B. Liu, “Learning to classify texts using positive and unlabeled data,” in Proc. 18th Int. Joint Conf. Artif. Intell., 2003, pp. 587–592.

[3] G. Chandrashekar and F. Sahin, “Asurvey on feature selection methods,” in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.

[4] Yuefeng Li, et al, “Relevance Feature Discovery for Text Mining” VOL. 27, NO. 6, JUNE 2015

[5] Y. Li, D. F. Hus, and S. M. Chung, “Combination of multiple feature selection methods for text categorization by using combinational fusion analysis and rank-score characteristic,” Int. J. Artif. Intell. Tools, vol. 22, no. 2, p. 1350001, 2013.

[6] Yuefeng Li, Yuefeng Li & Ning Zhong, “Mining Positive and Negative Patterns for Relevance Feature Discovery”.

[7] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, “Effective Pattern Discovery for Text Mining”

[8] C.Kanakalakshmi1, Dr.R.Manicka chezian, “Adaptive Relevance Feature Discovery for Text Mining with Simulated Annealing Approximation”.

[9] Harpreet Kaur, Rupinder Kaur, “Effective Pattern Discovery for Text Mining using Neural Network Approach”.

[10]Muthuvalli.A.R and Manikandan.M, “A Survey on Concept based Pattern Discovery for Text Mining”.

[11] A. Algarni and Y. Li, “Mining specific features for acquiring user information needs,” in Proc. Pacific Asia Knowl. Discovery Data Mining, 2013, pp. 532–543.