INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A BRIEF SURVEY OF VARIOUS APPROACHES FOR FEATURE OF TEXT MINING

Shabana Bee[1], Mr. Sumit Gupta[2]

M.Tech Student[1], Assistant Professor[2]
Department of Computer Science & Engineering, Lakshmi Narayan College of Excellence[1-2]
Email: Shabana2k10@gmail.com[1], sumitgupta888@gmail.com[2]

Abstract: - Text Mining, also known as knowledge discovery from text, and documents relating mining, refers to the process of extracting interesting patterns from very large text corpus for the purpose of discovering knowledge. Documents are used for clustering of text data documents .it is one of the new application of text mining where documents it is arranged as per different contents. This paper gives a brief description of various text mining approaches used by different researches which was based on term or pattern of the document. As in text mining there are basically two types of approaches one is term based approach and other is the phrase based .by application of text clustering for data mining to making it functionary Paper has explained the various approaches to text mining for finding the features of documents based on classification of text data. The Various evaluation parameters are also explained for comparing. The paper focused to find out the best approaches for better management of documents, for smart summarization of document and tried to justify approaches accuracy and feature.

Keywords: - Documents, Text data, Text Clustering, Feature, Text mining.

I. INTRODUCTION

Text Mining [1] is the discovery by computer of new previously unknown information by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar within web search. In search user is looking for something already known and has been written by someone else. Use of online information due to advancement of web technologies, text data mining approach

May be used for forming clusters of similar digital documents. It has become useful for extracting useful knowledge from web. Data mining technique is one of the best approaches for extracting useful information [2] the problem is pushing aside all material that currently is not relevant to your needs in order to find relevant information. In text mining, the goal is to discover unknown information, something that no one yet known and so could not have yet written down.

It is an interdisciplinary field involving information retrieval. Text understanding, Information extraction, Clustering, Categorization, Topic Tracking, Concept Linkage, Computational Linguistics, Visualization, Database Technology Machine Learning, and Data Mining .Text Mining offers a solution to this problem by replacing or supplementing the human reader with automatic system Undeterred by the text explosion. It involves analyzing of documents to discover previously unknown information. The information might be relationship or pattern that are buried in the document collection and which would otherwise be extremely difficult, if not possible, to discover text mining can be used to analyzed natural language documents about any subject, although much of the interest at present is coming from biological science. Originally, research in text categorization analyzed binary problem, where a document is either relevant or not. Text mining involves the application of technique from are as information retrieval, natural language processing, data mining and information extraction. Documents clustering has been studied in computer science literature and several technologies have been used for documents so, it has become necessary explain the basic feature of text minig [11].

Information retrieval (IR)) system identifies the documents in a collection which match a user's query. The most well-known IR system are search engine as Google, Which identify those documents on the world wide web that are relevant to a set of given words IR systems are often used in libraries, where the documents are typically not the book themselves but digital record containing information about the books. this is however changing with the advent of digital libraries, where the document being retrieved are digital version of books and journals IR system allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally intensive algorithm to large document collection, IR can speed up the analysis considerably by reducing documents for analysis. for e.g. if we interested in mining information only about protein interaction, we might restrict our analysis to documents that contains the name of a protein or some form of the web ‘to interact’ or one of its synonymous. So Documents features can contribute in differential to documents clustering .The term clustering denotes the process of arranging given objects in a group of similar objects together [3].

II. FEATURES OF TEXT MINING

1) Title feature

The word in sentence that also occurs in title gives high score. This is determined by counting the number of matches between the content word in a sentence and word in the title. In [4] calculate the score for this feature which is the ratio of number of words in the sentence that occur in the title over the number of words in the title.

2) Sentence Length

This feature is useful to filter out short sentence such as datelines and author names commonly found in the news articles the short sentences are not expected to belongs to the summary. In [5] use the length of sentence, which is the ratio of the number of words occurring in the sentence over the words occurring in the longest sentence of the documents.

3) Term Weight

The frequency the term occurrence with documents has been used for calculating the importance of sentence. The score of a sentence can be calculated as the sum of the score of words the sentences. The score of important score wi of word i can be calculated by traditional tf.idf method.

4) Sentence position

Whether it is the first 5 sentence in the paragraph, sentence position in text gives the importance of the sentences. These features can involve several items such as the position of the sentence in the documents, section and the

paragraph, etc., proposed the first sentence of highest ranking. The score for this feature in [6] consider the first 5 sentence in the paragraph.
this feature score is calculates with following sentence(5).

5) Sentence to sentence similarity
This feature is a similarity between sentences for each sentence S , the similarity between S and each other sentence is computed by the cosine similarity measure with a resulting value between 0 and 1 [6]. The term Weight wi and wj of term t to n term in sentences Si and Sj are represented as the vector. The similarity of each sentence pair is calculated based on similarity.

6) Proper Noun
the sentence that contains more proper noun (name entity is an important and is most probably include in the document summary . The score for this feature is calculate as the ratio of the number of proper noun that occur in the sentence,
over the sentence length.

S_f(6)S = No. Proper noun in S/Sentence Length (S)

7) Thematic Word
The number of thematic word in the sentence, this feature is important because term that occurred frequently in a document
are probably related to the topic. The number of thematic word
Indicates the word with maximum possible relativity. We used the top 10 most frequent content word for consideration as thematic. the score for this features is calculated as the ratio of the number of thematic words that occur in the sentence over the maximum summary of thematic word in the sentence.

S_f7(S) = No.thematic word in S/Max (No.thematic word)

III. CLASSIFICATION OF CONTENTIOUS NEWS ISSUES

Many topics in news may be considered controversial or have controversial parts or elements. This means that they include study of issues which people may strongly disagree about. For example, controversial issues may include exploring why some people in world are rich while other are poor, learning about the different sides of a conflict even whether or not a local authority should build coastal defense to protect land from erosion.

The coverage of contentious issues of a community is an essential function of journalism. Contentious issues continuously arise in various domains, such as politics, economics and environment; each issue involves diverse participants and their different complex arguments. However news articles are frequently biased and fail to fairly deliver conflict arguments of the issue. It is difficult for ordinary readers to analyze the conflict arguments and understand the contention they mostly perceive the issue passively, often through a single articles.
Advanced news delivery model are required to increase awareness on conflict news in this paper, this work present disputant relation based method for classifying news articles on contentious issues. We observe that the disputant of contentions, i.e., people who take position and participate in the contention such as Politian, companies, stack holders, civic groups, experts, commentators, etc., are an important features for understanding discourse. News producer primarily shape an articles on contention by selecting and covering disputants(Baker. 1994). Reader also intuitively understand the contention by identifying who the opposing disputant are the method help readers intuitively view the news articles through the opponents

based frame. It perform classification in an unsupervised manner: it dynamically identifies opposing disputants groups and classifies the articles according to their positions.

As such, it effectively help readers construct articles of contention and attains balanced understanding free from specific biased viewpoints.

Constructing Opposing Views

Online web forum discussing ideological and political hot-topics are popular. In this work, author are interested in dual sided oppose (there are two possible polarizing sides that the participants are take). For example, in a healthcare oppose, participants can take a for health-care instance or an again health-care stance. Participants generally pick a side (the websites provide a way for user to tag their stance) and post an arguments/justification supporting their instances.

The discourse of contentious issues in news articles shows different characteristics from that study in the sentiment classification tasks. First, the opponent of the contentious issue often discuss their different topics, as discussed in the example above. Research in mass communication has showed that opposing disputant talk across each other, not by dialogue, i.e., they martial different facts and interpretations rather than to give different answers to the same topics [13].

Second, the frame of arguments is not fixed as “positive versus negative.” We frequently observed both sides of a contention articulating negative arguments attracting each other. The forms of arguments are also complex and diverse to classify them as positive or negative; for example, an argument may just neglect the opponent's argument without positive or negative expressions, or emphasized a different discussion point. In addition, the position of a contention can be communicated without explicating expression of opinion or sentiments. It is often conveyed through objective sentences that include carefully selected facts. For example, a news articles can caused a negative light on a government program simply by covering the increase of deficit caused by it.

A number of recent works deals with debate stance reorganization, which is closely related task. They attempt to identify sides of a debate [9]. They deals with debate posts that are all on one coherent topic, for example, iPhone versus blackberry, and explicitly express arguments for or against to the topics, for example for an against iPhone or blackberry. Among these works work in [2, 10] is similar to our work as it does not assume a fixed classification frame nor required perceptively the discussed topic.

IV. PREPROCESSING

All words passes to preprocessing level. Irrelevant idioms are eliminated there. This process is also called as tokenization process. It consist of two kind of operations as stop list removals, stem word removals. [8]

1 Stop List Removals: It saves the system resources. Stop word has list of words. that are deemed or irrelevant and it is removing .It consists of articles(a, an, the), prepositions(for, in, at, etc.), and so on. A text documents is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text character by single white space. This tokenization representation than used for further processing. The set of different words are obtained by merging all text documents of a collection is called the dictionary of document collection.

Description: The description of the algorithm, we define first some terms and variables that will be frequently used in the following: Let D be the set of documents nd T= f (t1, t2, t3…tm) be the dictionary, i.e. the set of all different terms occurring in D.

Tokenization is the process of splitting a text stream into symbols, words, phrase or other meaningful elements called tokens. These tokens are used further text mining techniques. Word tokens are typically sent to preprocessing stages like stop-word removal and steeping, which are described later. They are also used input for feature extraction process. There are many ways of tokenization text stream into tokensA simple method would be just split text into blank space, but better method also take punctuations and other sings into consideration
The tokenization method used in this thesis would tokenize the following text string
"Hello! This is test number 11. It tests the words_punct-tokenizer!@ test66"
First by splitting it on blank space, then this is followed by splitting it on most special character. The tokenization string would then followed by tokens.
['Hello', '!', 'This', 'is', 'test', 'number', '11', '.', 'It', 'tests', 'the', 'word_punct', '-', 'tokenizer', '!@', 'test66']

2 Stem Word Removal: The group of different word may share the same word is called stems. For example drug, drugged, drugs, Different occurrence of the same word. Terms with a common stems would have same meaning. So it is filtering from the concern text documents. A stem is a natural group of words with equal (or very similar) meaning. After the steaming process every word is represented by its stems. A well-known rule based steaming algorithm
Has been originally proposed by porter [7]. He defines set of production rules to iteratively transform (English) words into their steaming algorithm every word is identified and the word co-occurrence are calculated with a score is calculated for each word.

Text Transformation

In order to analyze text data some separate presentation need to be developed for this that will evaluate data effectively these is such as Bags of words, Binary representation, TFIDF, etc. Here bag of word is collection of keyword for the categorization of word. This can be understood as the BOG= {‘India’, ‘Country’, ‘Production’} [10]. While in case of binary representation it shows that whether the word is present in the document or not, here if 0 represent the presence of word then 1 represent the word is not present in the document. TFIDF stands for term frequency keyword document frequency where TFIDF is the product of two statics term frequency and inverse document frequency. Various ways for determining the exact values of both statics exits. In the case of term frequency tf, the simplest choice is to use the *raw frequency of a term in the documents.* i.e. the number of time term *t* occurs in the document *d.*

The inverse document frequency is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total no of documents by the no of documents containing the term, and then taking the logarithm of that quotient idf.

Then TFIDF= tf*idf

Feature/Attribute Selection

Now selecting proper features from documents is necessary for evaluation of data, selection of representation come under this stage. here text in form of feature vector is organizing for data analysis in order to taken decision. As some training is required to discuss to learn different features so as per the training requirement features need to be select Classification.
On the basic of the obtained feature vector pattern are discovered the knowledge generation[12]. This can be understood as let a document have words then how many exactly matching the feature vector. On the basis of some lower limit of the matched words in the document one can classify that whether that document is relevant or not. Mostly this is done by some kind of system that undergoes some training, then testing.

Interpretation Evaluation

Once pattern are discovered then the compared result of the system are need to be evaluate that either result obtained is correct or not. If the obtained result is not correct then the training parameter need to be change, if the result are still vary then the pattern are discovered are also need to be re-shuffled or change as per requirement.

Precision=true positive/ (true positive + false positives)

Recall=true positives/ (true positive + false negative)

F-score= 2*Precision*Recall/ (Precision + Recall)

In order to evaluate result there are many parameter such as accuracy, precision, recall, F-score, etc. obtaining values can be put into the mentioned formula to get better result.

V. TECHNIQUE EXPLANATION AND COMPARISION

KNN (K Nearest Neighbors algorithm) in is used which utilize nearest neighbor property among the items. This algorithm is easy to implement with high validity and required no prior training parameters. Although K nearest neighbor is also identified as instance based learning in other words classification of items is quite slow. In this classification techniques distance between the K cluster center and classifying item is calculated then assign item to cluster having minimum distance from the cluster center. In case of text mining features from the document is extracted then k labeled node is select randomly which are supposed to be cluster center and rest of nodes or document are unlabeled nodes. Finally distance between labeled and unlabeled node is calculate on the base of feature vector similarity. In this algorithm distance between nodes are estimate in log(k) time **Support Vector Machine** (SVM) in is quite famous soft computing technique for item classification which is based on the input feature vector quality and training of the support vector machine. In this technique a hyperplane is built between the items this hyperplane classify the items into binary or multi class. In order to find the hyperplane equation is written as P = B+XxW where X ia an item to be classify then W is vector while B is constant. Here W and B is obtained by the training of SVM. So SVM can perfectly classify binary items by using that calculated hyperplane.

Fuzzy classification in, has classify image data which is highly complex and required stochastic relations for the creation of feature vector from images. Here different types of relations are combined where members of the feature vector is fuzzy in nature. So this relation based image classification is highly depending on the type of image format as well as on the threshold selection.

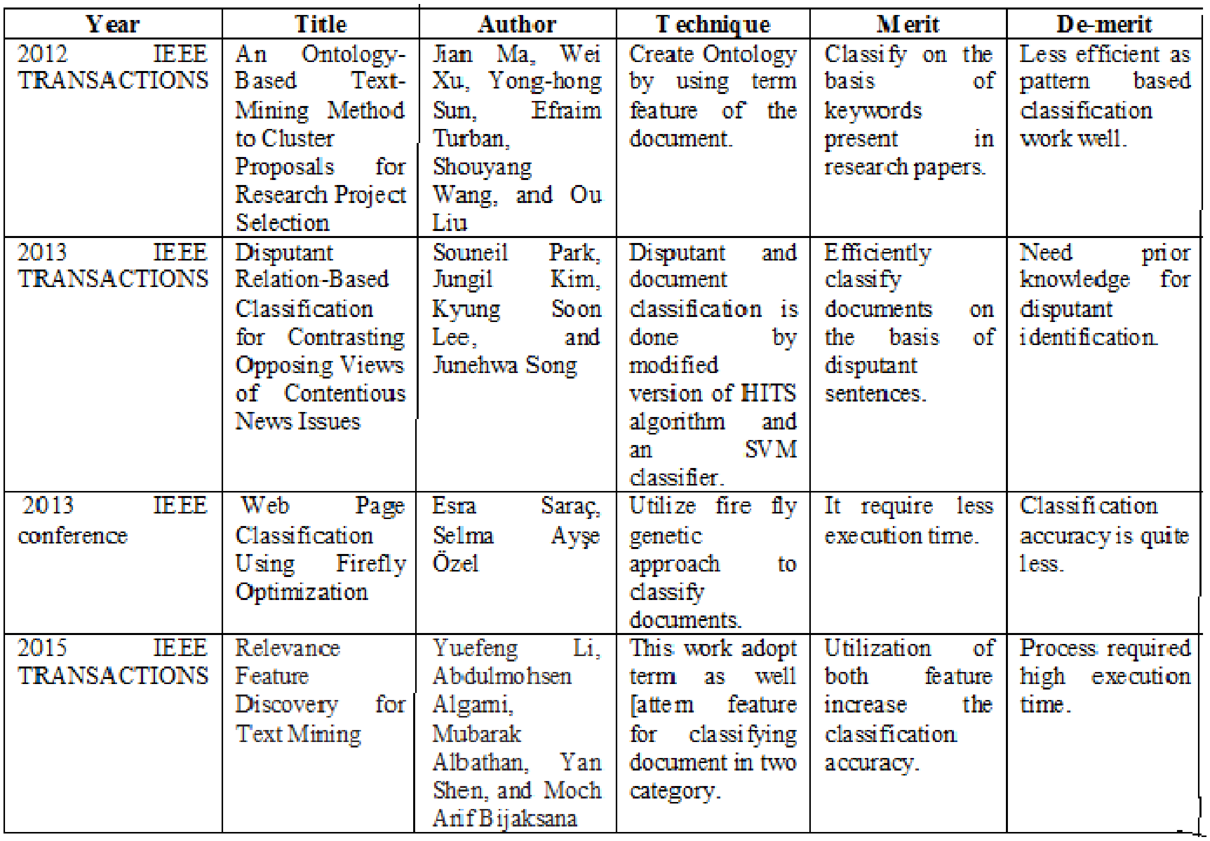
Fire Fly

This is a genetic approach by which classification of data was done. In this technique probable cluster centers are randomly collect which is term as chromosome. Here classification items are fire fly and distance between those is calculated based on the light intensity. So probable solution which has minimum distance from other items in the cluster is considered as the final solution.

Ontology Based Classification

In this work a dictionary is maintained which is a collection of strong or keywords from the particular cluster. So here a supervised learning is done. Whole work of classification is done by finding similarity between the document features with the ontology. Ontology is not only explaining similarity of documents feature but also repository of concepts and terms, its relation in their documents and text mining [12]

Comparison Table



Year	Title	Author	Technique	Merit	De-merit
2012 IEEE TRANSACTIONS	An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection	Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu	Create Ontology by using term feature of the document.	Classify on the basis of keywords present in research papers.	Less efficient as pattern based classification work well.
2013 IEEE TRANSACTIONS	Disputant Relation-Based Classification for Contrasting Opposing Views of Contentious News Issues	Souneil Park, Jungil Kim, Kyung Soon Lee, and Junehwa Song	Disputant and document classification is done by modified version of HITS algorithm and an SVM classifier.	Efficiently classify documents on the basis of disputant sentences.	Need prior knowledge for disputant identification.
2013 IEEE conference	Web Page Classification Using Firefly Optimization	Esra Saraç, Selma Ayşe Özel	Utilize fire fly genetic approach to classify documents.	It require less execution time.	Classification accuracy is quite less.
2015 IEEE TRANSACTIONS	Relevance Feature Discovery for Text Mining	Yuefeng Li, Abdulmohsen Algami, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana	This work adopt term as well [attem feature for classifying document in two category.	Utilization of both feature increase the classification accuracy.	Process required high execution time.

VI. CONCLUSION

As the writing work of different articles from laboratory, organization, press media, institutes are increasing day by day then publishing their work is also increase which is done by most of the journals, newspaper, organizations .Here survey paper has covered an important issue of document classification. Various approaches with their required feature are discussed in details. In this paper one concept of disputant is explain with a mix of keywords of that documents, this term has given an effective result that are highly dynamic , as it is acceptable for paper of different field.

REFERENCES

[1] Selma Ayşe Özel. Esra Saraç “Web Page Classification Using Firefly Optimization “, 978-1-4799-0661-1/13/$31.00 © 2013 IEEE.

[2] S. Somasundaran And J. Wiebe, “Recognizing Stances In Ideological Online Debates,” Proc. Naacl Hlt Workshop Computational Approaches Analysis And Generation Emotion In Text (Caaget ’10), Pp. 116-124, 2010.

[3] G. Salton, C. Buckley, “Term-Weighting Approaches In Automatic Text Retrieval” Information Processing and Management 24, 2008. 513-523.

[4] L. Suanmali, N. Salim, M.S. Binwahlan, “Srl-Gsm: A Hybrid Approach Based On Semantic Role Labeling and General Statistic Method For Text Summarization”, Research Article- Journal Of Applied Science, 2010.

[5] M. K. Dalal, M. A. Zaveri, “Semi-supervised Learning Based Opinion Summarization And Classification For Online Product Reviews”, Hindawi Publishing Corporation Applied Computational Intelligence And Soft Computing, Volume 2013.

[6] A. Kiani, M. R. Akbarzadeh, “Automatic Text Summarization Using: Hybrid Fuzzy Ga-Gp”, International Conference On Fuzzy Systems, 2006 Ieee. [11] Base Paper

[7] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu “Effective Pattern Discovery For Text Mining”. Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012

[8] Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu. “An Ontology-Based Text-Mining Method To Cluster Proposals For Research Project Selection”. Ieee Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012

[9] Disputant Relation-Based Classification For Contrasting Opposing Views Of Contentious News Issues Souneil Park, Jungil Kim, Kyung Soon Lee, And Junehwa Song. IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013.

[10] S. Park, K.S. Lee, And J. Song, “Contrasting Opposing Views Of Contentious Issues,” Proc. 49th Ann. Meeting Assoc. Computational Linguistics (Acl ’11), Pp. 340-349, 2011.

[11] Twinkle Svadas, Jasmine Jha “A Literature Survey On Text Document Clustering and Ontology based techniques” Published in International Journal of Innovative and Emerging Research in Engineering ,Volume 1,Issue 2,2014

[12] Prof. S.R. Durugkar ,Madhuri Malode “Survey Paper on Clustering of Documents Based on Partitioning the Feature ”Published in International Journal of Engineering Research and Technology(IJERT), Volume 3,,Issue 1 ,2014

[13] R.Jensi and Dr.G.Wiselin Jiji “A Survey On Optimization Approaches To Text Document Clustering ”Published in International Journal on Computational and Application (IJCSA) Vol.3,No.6,Dec.2013