International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-1 Issue-6, August 2012

A Cogitate Study on Text Mining

Y. Jahnavi, Y. Radhika

***Abstract*: The vast amount of digitalized textual content has given rise to the need for sophisticated text mining techniques. Text Mining is the process of analyzing and extracting the useful information from a set of semi structured and unstructured documents by applying machine learning and natural language processing techniques. It is easy for the people to assimilate from the categorized text documents. Even though a large research has evolved into this problem, there is a survey that indicated trends and directions. In this paper a cogitate study on preprocessing, term weighting algorithms, concept based term weighting algorithms, pattern discovery, categorization, domain ontology based frame work for text mining and summarization techniques is presented. In addition, a number of successful applications of text mining are discussed**.

***Index Terms-Classifiers, Term Weighting, Text Mining*.**

I. INTRODUCTION

Text Mining refers to the discovery of non-trivial, previously unknown and potentially useful information from a large collection of textual documents. [1][3]. Text Mining is most important since now a day, most of the information is available in textual form. The vision of text mining is as data mining on unstructured data. Text Mining takes the unstructured text documents as the initial input and most of the approaches in text mining have been applied to extract structured data sets, called intermediate forms. Different intermediate repress- entations of text areWord, Phrase, Pattern, Concept, Paragraph and Document. Any usual data mining techniques can be applied on these intermediate forms.Text categorization is the task of automatically sorting a set of documents into categories such as topics. As the scenario has changed from the traditional static text set to dynamic text stream, where a text stream is a sequence of chronologically ordered documents, the topic extraction model should not only focus on terms statistical information such as frequencybut also on concepts[2]**.**

Automatic topic extraction in stories is the process ofidentifying important terms in the story that satisfy topicality property , where in news documents is the process of identify hot terms that satisfy both topicality and pervasiveness property[4]**.**Many applications such as business management and market analysis and many more can benefit by the use of the information and knowledge extracted and categorized from a large amount of textual data. However, searching for useful and interesting patterns was still an open problem. In this paper we present different techniques for extracting useful information. We also discussed the usage of extracted patterns and applications of text mining.

II. PREPROCESSING

For preprocessing the documents, Hotho and Stumme (2003) used the Text Mining System[1][5]. They propounded, let us first consider documents to be the collection of terms. Let $tf(d,t)$ be the absolute frequency of term $t \in T$ in document $d \in D$, where D the set of documents and $T = \{t_1, t_2, \ldots\ldots\ldots, t_m\}$ the set of all different terms occurring in the document D. The term vectors are denoted by $\vec{t_d} = \left(tf(d,t_1), \ldots\ldots tf(d,t_m)\right)$ .Later on they have adopted a slightly more advanced approach which consists of stop word removal, stemming of terms, and pruning of terms that appear infrequently. In the sequel, they have taken tfidfalgorithm into account, which modify the term vectors $\vec{t_d}$ accordingly [5].

Stop list Elimination:

The objective of stop list elimination is to save system resources. Stop list are the list of words which are considered as non–evocative and to be filtered out prior to or after processing text. They typically comprise prepositions, articles, and so on. There is no specific list of stop words for all applications and these stop words are controlled by the human but not automated.

Stemming:

In the next step, the text documents have to be processed using the Porter stemmer. Removing suffixesis particularly useful in the field of text mining and information retrieval. Instead of using the original terms in the documents, stemmed terms are used to construct a vector representation $\vec{t_d}$ for each text document. Thelength of the consequential vectors after stemming is given by the number of different stemmed terms in the text corpus [5]. Text is a collection of documents and each document is composed of a set of words or terms. Terms with a common stem would generally have similar meaning. These words," test, testing, tester, tests" have same stem called test. The performance of any search system is improved if all these word are conflated into a single word called test. The choice to perform stemming is a tradeoff between precision of a search and the overhead diminution [6].In two ways the suffixes can be stripped i.e., either by using stem dictionary or by using suffix list. Porter stemmer algorithm is implemented by using suffix list. The main merits of the present porter stemmer algorithm are that it is speedy small, and reasonably simple. Porter stemmer algorithm is implemented as a simple direct suffix stripping approach but not a linguistic exercise. This popular algorithm may not work for all words, for example: it seems unreasonable to conflate the words 'RELATE' and 'RELATIVITY' to a single word.

Manuscript received on August, 2012.
Y. Jahnavi, Computer Science & Engineering Department, Research Scholar, GITAM University, India.
Dr. Y. Radhika, Computer Science & Engineering Department Associate Professor, GITAM University, India.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

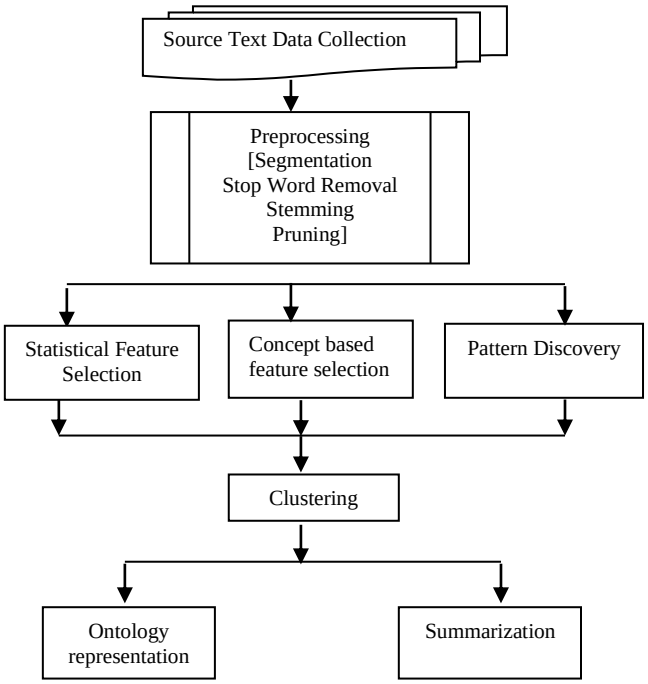

Fig .1 Overall System Architecture.

Pruning:

All infrequentterms have to be leftover for experimental evaluations. For a pre-defined threshold δ , a term t is discarded from the representation (i. e., from the set T),

$$if \sum_{d \in D} tf(d,t) \le \delta$$

The basis behind pruning is with the aim of infrequent terms doing not be usefulin identifying appropriate clusters.While stemming, pruning and term weighting was performed, they have always performed them in the order in which it have been scheduled here [5]. The overall system architecture is shown in fig. 1.

III. TERM WEIGHTING ALGORITHMS

The first important step in Automatic Topic Extraction is the process of identifying important terms & sentences. The basic idea is to assign weights to terms and sentences based on their frequency and some other statistical information. Terms with high weights capture the topical information. Some algorithms exist which extracts terms satisfying the property of pervasiveness. There exist some algorithms which are simple and easy without taking semantic relationship into consideration. But the performance of such algorithms is low. The importance of a term is calculated based on the semantic structure of each term, rather than the frequency of a term. Recently a new concept based modelwas presented to bridge the gap between Natural Language Processing and text mining [7][8]. To overcome the problems of synonymy and polysemy problems in term based approaches, pattern-based approaches are introduced.

(i). TFIDF Weighting:

The tf–idf (term frequency–inverse document frequency) is a weight assigned to the terms in Text Mining and Information Retrieval. This weight is a statistical measure used to assess the importance of aword in a collection of documentsor corpus. The significance of the term increases with respect to the number of times a word appears in the document[9].

Tf-idf weighs the frequency of a term in a document with a factor that discounts its significance when it appears in almost all documents. Therefore terms that appear too rarely or too frequently are ranked lower than terms that balance between the two boundaries and, hence, are likely to be better able to contribute to clustering results [10].

Definition: The tfidf (term frequency-inverted document frequency) of term t in document d is defined by:

$$tfidf(d,t) = \log(tf(d,t)+1) * \log(|D|/df(t))$$

$tfidf$ where $df(t)$ is the document frequency of term t that counts in how many documents the term t appears. If $tfidf$ weighting is applied then they replace the termVectors $\vec{t_d} = (tf(d,t_1),.....tf(d,t_m))$

$$by\ \vec{t_d} = (tfidf(d,t_1),.....tfidf(d,t_m))$$

Mathematical Details:

The term count in the given document is basically the number of times a given term appears in that document. This count is generally normalized to prevent a bias towards longer documents (which may have a higher term count despite of the actual importance of that term in the document) to give a measure of the importance of the term t_i within the particular document d_j . Thus the term frequency could be defined as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Where $n_{i,j}$ is the number of occurrences of the considered term in document d_j , and the denominator is the sum of number of occurrences of all terms in document d_j .

The inverse document frequency is a measure of the general importance of the term (obtained by dividing the number of all documents by the number of documents containing the term, and then taking the logarithm of that quotient). This could be represented as:

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

With

- $|D|$: total number of documents in the corpus
- $|\{d : t_i \in d\}|$: Number of documents, where the term t_i appears (that is $n_{i,j} \neq 0$).

Then

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i$$

A high weight in tf-idf is assigned by a high term frequency in the particular documentand a low document frequency of the term in the entire collection of documents; the weights hence tend to filter out most common terms. In [11]tf*idf weighting scheme is used for text representation. In addition to TF-IDF, the global IDF and entropy weighting scheme is proposed in [12] and improves performance by an average of 30 percent. Various weighting schemes were proposed in [13][14][15].

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

(ii) *Novel Concept of TF*PDF:*

Bun andIshizuka proposed system works on the basic concept that whenever there is a hot topic in the air, the topic would be discussed frequently in many news documents from major newswire sources[10]. Rather than calculating the term weights from the entire mixed corpus, they calculated from each channel.The terms that explain the hot topics, should appear frequently in many documents in each channel which would be weighted significantly. The novel concept of TF*PDF is that a term is heavily weighted whenever it is discussed in most of the channels and less weighted if it occurs frequently in a few channels and disappears in all other channels**.**

TF*PDF is introduced to accomplish the novel concept which is different from the usual TF-IDF algorithm [10].In TF*PDF algorithm, the weight of a term from a channel is linearly proportional to the terms within channel frequency, and exponentially proportional to the ratio of the document containing the term in the channel. The total weight of a term would be the summation of term's weight from each channel as follows:

$$W_j = \sum_{c=1}^{c=D} |F_{jc}| \exp\left(n_{jc}/N_c\right)$$

Where

Normalized term frequency $|F_{jc}| = F_{jc} \Big/ sqrt\left(\sum_{k=1}^{k=K} F_{kc}^2\right)$

W_j =Weight of term j;

F_{jc} =Frequency of term j in channel c;

n_{jc} =Number of documents in channel c, where term j occurs;

N_c =Total number of documents in channel c;

K =Total number of terms in a channel;

D =number of channels [10]

(iii) *Term weight calculation based on the life cycle model*

The hot topics in the news reports are completely depend upon the time variation,whereas traditional term weighting approaches are insufficient to extract hot terms based on the variation in the usage of terms over time. The authors, Chen, Luesak and Chou (2007), proposed a new method which considers variations in the frequency that terms are used over time in turn to correctlyrecognize hot topics[16].

Hot terms have the following properties:

- Pervasiveness: This property refers to the frequency with which a term appears in a set of documents. The higher the frequency is, the more pervasive a term becomes.
- Topicality: This property refers to the variation in the frequency of usage of a term over time. A term is more topical if its usage varies greatly.

The weight of a hot term is defined as the sum of the weights given by each of these two properties. The procedure of hot term extraction is shown in fig.2.

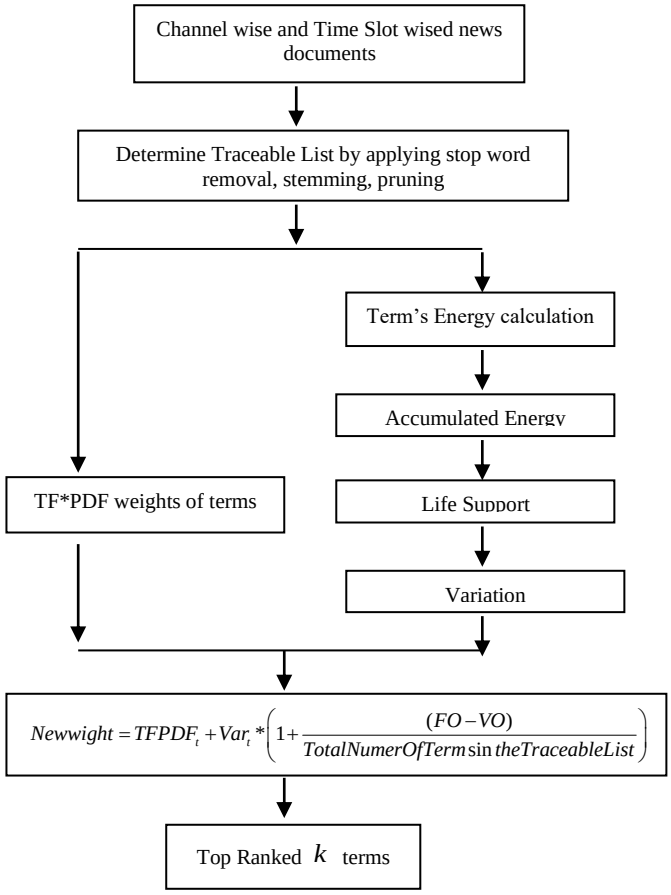

Fig.2 Term weight calculation based on lifecycle model

The former property is measured by TF*PDF[10], which has been used for topic detection in news documents.

The later property is calculated by tracking the lifecycle of the term. According to the life cycle model , the accumulated energy $E_{t,s}$ of term t measures the frequency of term t appearing in a specified time slot s. The life support value of t at time slot s is calculated as the logarithm of $E_{t,s}$ represented as $lifesupport_{t,s}$. The variation of its usage over time can be computed as :

$$V_t = \sqrt{\frac{1}{N}\sum\left(life\sup port_{t,s} - \overline{life\sup port}\right)^2}$$

Where N is the number of time slots

$life\sup port$ is theaverage of life support value for term t.

The overall weight of term t is measured by combining the two characteristic together.

$$Newwight = TFPDF_t + Var_t * \left(1 + \frac{(FO - VO)}{TotalNumerOfTerm \sin theTraceableList}\right)$$

By sorting the terms in the candidate list with the combined weight, the top ranked k terms can be chosen as hot terms. These hot terms reflects the hot topics.

(iv) ThePosition Weighted TF*PDF

There are most commonly used term-weighting algorithms such as TF*IDF and their improvement-Incremental TF-IDF[17][18][19][20][21]. TF*PDF is different from TF–IDF, where it assigns higher weights to the terms that occur in many documents in many channels. At all times the news report consists of title and the body.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The terms in the title are more important and reflect the main idea of the news document than the terms in body text [22].

The position score to term j in document k in channel i is defined as:

$$ps_{ik}(j) = \begin{cases} 3 & j \in \text{the title of the k}^{th}\text{document in channel i} \\ 1 & j \notin \text{the title of the k}^{th}\text{document in channel i} \end{cases}$$

The total position weight of a term will be the summation of average weight from each channel as follows:

$$pw(j) = \sum_{i=1}^{|C|} \sum_{k=1}^{|N_c|} ps_{ik}(j) / N_{jc}$$

By considering the effect of the term's position, a new term weighting scheme was proposed by combining TF*PDF with position weight. This algorithm is called position weighted TF*PDF

$weight_j = W_j * pw(j)$ Where W_j is the TF*PDF weight.

The top ranked k term can be chosen by sorting the terms with the modified TF*PDF i.e., position weighted TF*PDF.

(v) Combining Term Frequency with Information Theory Functions or Statistic metrics

There exist different unsupervised term weighting algorithms such as binary tf , $tf.idf$ and its variants and unsupervised term weighting schemes such as $tf.\chi^2$, $tf.ig$ (information gain), $tf.gr$ (gain ration), $tf.OR$ (Odds Ration) etc. Supervised term weighting approaches outperformed the unsupervised traditional term weighting methods. Positive category terms contribution is more for separating the positive samples from the negative samples. i.e., the terms which are in positive category should be weighted more than the terms in negative category.

The contingency table for Category c_i and Termt$_k$

	t_k	$\overline{t_k}$
Positive Category: c_i	a	b
Negative Category: $\overline{c_i}$	c	d

All the existing algorithms cannot distinguish the terms whose a: c ratios are same.

The famous idf factor is computed as $\log\left(\frac{N}{a+c}\right)$,Where N= a+b+c+d.

In this tf.idf algorithm the idf factors are same for any a:c values . To overcome this problem a new supervised term weighting algorithm $tf.rf$ (Relevance Frequency) was proposed[23]. On the converse, by adopting the rf scheme, each term is assigned more suitable weights in terms of dissimilar categories.

The relevance frequency is expressed as :

$$rf = \log\left(2 + \frac{a}{c}\right)$$

The constant value 2 in the rf formula is assigned because the base of this logarithmic operation is 2.

In any case if c=0, then the divide by zero error occurs.

To avoid such situation, the minimal denominator is set to 1.

Thus rf expressed as:

$$rf = \log\left(2 + \frac{a}{\max(1,c)}\right)$$

When combining with tf , the final $tf.rf$ expressed as:

$$tf.rf = tf * \log\left(2 + \frac{a}{\max(1,c)}\right)$$

This algorithm considers the frequency of only relevance documents.The rf does not consider the values of d, because d will lessen the significance of a and c.

(vi) TF-ICF (Term Frequency and Inverse Corpus Frequency)

TF-IDF can able to find weights only for static data streams. But the static methods cannot satisfy all these problems [24].

(i) The Document Frequency distribution of a smaller data set cannot approximate a larger data set.

(ii) The Document Frequency of one corpus cannot approximate to another corpus.

(iii) Composing a corpus such that it covers all the words commonly found in writing.

A new term-weighting scheme called TF-ICF in which the weight of each term is calculated as follows:

$$w_{ij} = \log(1 + f_{ij}) \times \log\left(\frac{N+1}{n_j+1}\right)$$

Where f_{ij} is the number of occurrences of term j in document i .

n_j is the number of documents in the corpus where term j occurred one or more times.

N is the total number of documents.

TF-ICF can produce clusters that are comparable quality to those generated by the widely recognized term weighting schemes, such as TF-IDF[24].

(vii) Statistical Property (Term Distribution) based Term Weight Algorithm

Since term frequency is not the only discriminator,there are other statistical characteristics of terms used for discrimination i.e., term distribution.A term with high frequency and close to hypo-dispersion distribution should be given higher weight than one with lower frequency and close to intensive distribution. Based on this characteristic, a term distribution based term weight algorithm is proposed[25]. Since TF-IDF only takes term frequency into consideration, it suffers from the following drawback.TF calculates term weight based on their frequency .i.e., term weight is positive correlated to their frequency. Actually, term with higher frequency may only be intensively distributed in a part of the document. Such terms are used to represent the content of the part instead of the whole document. However TF algorithm will assign a higher term weight to such terms. It is inadequate to only consider term frequency to assign weights to the terms.Distribution based of term weight algorithm assigns weights based on Uniform Distribution Extent and Spread Extension.

(viii) Concept –Based Mining Models

Statistical approaches obviate the need to analyze the conceptual relevance of terms or phrases in a document. In text mining, selecting important features which contribute to the formation of proper clustering algorithms is important.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

In text mining techniques, the term frequency of a term reflects the importance of the term in the document[26]. On the other hand, two terms can have the same frequency in their documents;however one term contributes more to the significance of its sentences than the other term. The importance of a term is calculated based on the semantic structure of each term, rather than the frequency of a term. Recently a new concept based model[28][29] was presented to bridge the gap between Natural Language Processing and text mining. In[27], the authors Shehata, Karray, and Kamel proposed mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis and concept based similarity measure.

Initially each document should be zonedand well defined sentence boundaries are identified. The terms in the sentences are identified are labeled based on the PropBank notations [31]. Then the sentences in the document arerepresented in the form of verb argument structure. A sentence may have more than one verb argument structures. This arrangement allows the creation of a compound meaning representation from the meaning of the individual concepts in a sentence.

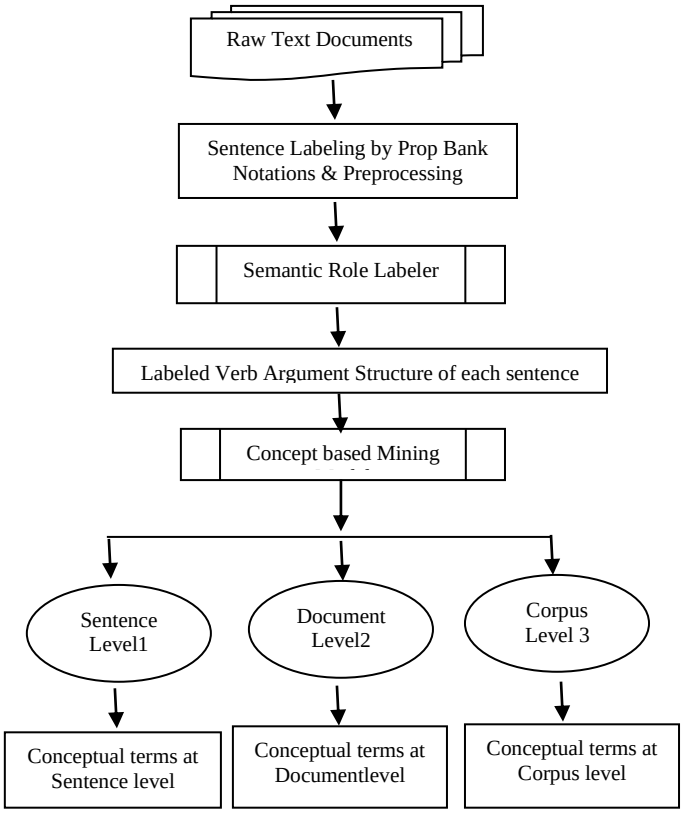

Fig. 3 Conceptual mining

The verb argument structure of the sentences is used to find the conceptual terms. The study of the roles associated with verbs is referred to as a thematic role or case role analysis[30]. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence. The advantage of concept based model is that it can effectively discriminate between meaningful terms which don't describe sentence meaning.Conceptual Term Frequency (CTF) in Sentence s is the number of occurrences of concept c in verb argument structures of sentence s.Conceptual Term Frequency (CTF) in Document d is the average of the ctf values in its sentences of document d.CTF in corpus level is a global measure calculated based on ctf in sentence and document.The overall process of conceptual mining is represented in fig. 3. These conceptual measures accurately influence the clustering process.

(ix).Pattern Based Approaches:

Most of the feature extraction methods have adopted the fundamental term based approaches. But they suffer from the problems of polysemy and synonymy. There has been a large amount of work in the field of Information Retrieval based on term based methods because of its simplicity and their computational efficiencies [32][33][34][35][36]. Though the term-based methods suffer from the problems of polysemy and synonymy, the advantages include efficient computational performances as well as mature theories for term weighting.The sequential patterns have gained interests in text mining community from the past several years. The reasons of using sequential patterns approach over term based approaches in text mining community are

i. As compared to term based on single words, a feature based on sequence pattern captures temporal relationships between words and phrases.
ii. A language model based on sequential patterns has more expressive power than those approaches based on words.

IV. STATISTICAL APPROACHES NAÏVE BAYES CLASSIFIER

A naïve Bayes classifierappliesBayesian statistics on the features that make the classification process based on the Bayes theorem.

By Bayes Theorem

$$P(C_i/X) = \frac{P(X/C_i)P(C_i)}{P(X)}$$

Essentially, the presence or absence of a exacting feature of a class is assumed to be disparate to the presence or absence of any other feature [37] [38].During its training phase, a naïve Bayes classifier learns the posterior word probabilities. The main strength of naïve Bayes algorithm lies in its simplicity. Since the variables are mutually independent, only the variances of individual class variables need to be determined rather than handling the entire set of covariance. It is improving its accurateness constantly when identifying wrong classifications by considering to each user's preferences thus allowing continuous rectified training of the model.

DECISION TREE

A decision tree is a hierarchical predictive model for supervised learning that expands a tree of decisions and their possible consequences. The outcomes can be discrete or continuous. Each leaf represents a unique classification and branches represent the test condition that lead to the classification at various leaves. Popular decision tree based learning methods are CART, ID3, C4.5 and Naïve Tree[39].A decision tree for uncertain data handling is also introduced in paper [40].

SUPPORT VECTOR MACHINE (SVM)

An SVM is a supervised learning method that subjects every category to a separate binary classifier. The instances should be some distance away from the hyper plane for better generalization.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The optimal separating hyper plane is the one which maximizes the margin between positive and negative examples of training data. If the problem is nonlinear then those nonlinear instances will be mapped to a new space by a nonlinear transformation with a suitable basis function. By the suitable kernel functions, it can learn polynomial classifiers, radial basis functions etc and thus acquiring universal learning ability [37] [38].

K-NN

KNN algorithm is simple and efficient lazy classifier. Initially all the training tuples are simply stored and when the test tuples are given it classify them based on the similarity (Euclidean measure) to the stored training tuples. The performance of KNN is determined by the choice of K. The value of K can be determined experimentally. Choosing K is difficult when their exist noise in the data. The main strength of KNN classifier is that it provides accurate results when the low noise training data is given as input and its learning phase is fastbecause it is lazy learner. Its performance is degraded because of its inefficiency in the case of high dimensional and large scale data sets.

All data mining categorization techniques such as hierarchical clustering, partitioned clustering, density based algorithms, self- organizing map algorithm etc are also applicable for categorizing unstructured data.

V.DOMAIN ADAPTATION MODELS

Labeled instances are required for text classification in order to train a high quality model. It is not that much easy to label all the instances in a problem domain. To solve this problem, domain adaptation techniques are introduced [41][42]. Domain adaptation techniques capture the shared knowledge from other related domains where labeled data are available. Domain adaptation techniques can be classified into two categories, Instance based approaches [43] and Feature based approaches [44]. Instance-based approaches try to find some reweighting on the source data, such that the gap between source distribution and target distribution is reduced. Feature-based methods try to discover a shared feature space on which the distributions of source and target domains are brought nearer. When the information gap between source domain and target domain are large then these domain adaptation techniques cannot work. To overcome this problem Xiang, Cao, Hu, Yang and Fellow introduce a bridge between the different domains by barrowing additional knowledge sources that are available such as Wikipedia [45]. Here semi supervised learning is applied for domain adaptation problems based on the auxiliary data.

VI. MULTIDIMENSIONAL CLUSTER MODELING

There are useful language applications such as EuroNet, WordNet, and Part-of-speech tagging, etc. that can enhance text mining performances [13].Instead of the traditional one vector model, the multidimensional vector model considers more Diverse criteria for identifying the similarity

Concept vectors (CVs): The CVs are derived from the background knowledge. The Word Net is one of the sources of background knowledge on words[46]. There are different types of conceptual vectors Direct CV (DCV), Kind of Vector (KV), and Part of Vector (PV) etc,

- DCV: Each element contains the synonyms of w obtained from WordNet.
- KV: Each element contains the hypernyms of w taken from WordNet.
- PV: Each element contains the regular homonyms of w listed on WordNet.

WordNet

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: To produce a combination of Dictionary and Thesaurus. As of 2006, the database contains about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs; in compressed form, it is about 12 megabytes in size. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. Every synset contains a group of synonymous words or collocations (a *collocation* is a sequence of words that go together to form a specific meaning, such as "car pool"); different senses of a word are in different synsets.

Most synsets are connected to other synsets via a number of semantic relations. These relations vary based on the type of word, and include:

- Nouns
  - hypernyms: Y is a hypernym of X if every X is a (kind of) Y (canine is a hypernym of dog)
  - hyponyms: Y is a hyponym of X if every Y is a (kind of) X (dog is a hyponym of canine)
  - coordinate terms: Y is a coordinate term of X if X and Y share a hypernym (wolf is a coordinate term of dog, and dog is a coordinate term of wolf)
  - holonym: Y is a holonym of X if X is a part of Y (building is a holonym of window)
  - meronym: Y is a meronym of X if Y is a part of X (window is a meronym of building)
- Verbs
  - hypernym: the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (to perceive is an hypernym of to listen)
  - troponym: the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (to lisp is a troponym of to talk)
  - entailment: the verb Y is entailed by X if by doing X you must be doing Y (to sleep is entailed by to snore)
  - coordinate terms: those verbs sharing a common hypernym (to lisp and to yell) [46]

VII. DOMAIN ONTOLOGY BASED RAMEWORK FOR TEXT MINING

The semantic information cannot be represented with the traditional text mining algorithms**.** Thus high accuracy cannot be achieved. Ontology provides supportfor semantic information representation.Ontology stores the main concepts and their inter-relations and therefore providing a collective knowledge.

Ontology initially started in the field of philosophy, and later, it is now gaining a specific role in Artificial Intelligence.Gruber proposed the most popular definition in 1993, “Ontology is a conceptual model of a clear specification“[47].Different Authors have proposed different definitions.In 1998, Studer deeply analyzed all the models and proposed that, “Ontology is the explicit formal specification of a shared conceptual model” [48].

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The knowledge expressed by ontology should be accepted by the entire group not by the single group. Moreover, it should reflect the common recognized knowledge of the end users.According to the dependency on the specific fields, ontology can be subdivided into four categories: General Ontology, Domain Ontology, Task Ontology and Application Ontology [49].To build domain ontology, the domain knowledge should be fully understand, domain core concepts and their inter-relations should be identified. From this knowledge the domain ontology is build. If the constructed domain ontology doesn't have available thesaurus, then it is necessary that the experts should organize and construct them, to describe the core concepts and their inter-relation. There exist different ontology languages to describe the domain knowledge ontology.

VIII. ADVANTAGES OF DOMAIN ONTOLOGY BASED FRAMEWORK FOR TEXT MINING.

1) Compared to the traditional model, this model mines concepts and relations more accurately.
2) The traditional text mining generates text database with low-level form. High-level rules are considered as the summary of low level rules. These high level rules are nothing but the concepts [50][51].

The concepts and their inter relations represented by the ontology are useful for comparing text documents to form clusters [52].However, the study of text mining based on domain ontology is still in the starting stage, and has not yet formed anorganized study.

IX. SUMMARIZATION TECHNIQUES

Automatic summarization is the creation of shortened version of a text by a computer program [53] [54]. Summarization is a reductive transformation of source text to summary text through context reduction. Summarization techniques are used in search engines such as Google.

Types of Summarization techniques:

- Extractive summarization It copies the most important information by the system to the summary. i.e., it assigns scores to the sentences and paragraphs of the documents and extracting those with highest scores.
- Abstractive Summarization These methods built an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original.

Abstractive methods are quite weak, so most research has focused on extractive summarization techniques.

Key phase extraction is the first step before summarizing using extractive summarization technique.

- Text Rank VsLex Rank Text Rank uses for single-document summarization. LexRank has been applied for multi-document summarization.
  In lexRank, when summarizing multiple documents, there is a great risk of selecting duplicate or highly redundant sentences to place in the same summary. To address this issue, lexRank applies a heuristic post-processing step that builds up a summary by adding sentences in rank order, but discards any sentences that are too similar to ones already placed in the summary.
- Difficulties in evaluating summaries The most common way of in formativeness of automatic summaries is to compare them with human-made model summaries. However, as content selection is not a deterministic problem, different people would choose different sentences, and even the same person may choose different sentences at different times.
  Besides human variability, the semantic equivalence is another problem, because two distinct sentences can express the same meaning but not using the same words. This phenomenon is known as paraphrase.

X. TEXT MINING APPLICATIONS

The main Text Mining applications [3] are most often used in the following sectors:

- Publishing and media.
- Telecommunications, energy and other services industries.
- Information technology sector and Internet.
- Banks, insurance and financial markets.
- Political institutions, political analysts, public administrationand legal documents.
- Pharmaceutical and research companies and healthcare.

XI. CONCLUSIONS

This paper presented preprocessing techniques as well as different categorization techniques. We also discussed the strengths and weaknesses of each technique. From the survey presented we observe that significant work has been done in the field of preprocessing, term weighting andtext categorization techniques. These indeed can be applied to a wide range of web applications. The direction now points to context-based approaches and pattern based approaches. The study concludes that the concept based analysis bridges the gap between Natural Language Processing and Text Mining. On the other hand the algorithms and strategies of text mining have to be improved. There is yet more scope for future research in the field of context-based text classification.

REFERENCES

[1] S. M. **INDURKHYA**, N. **ZHANG**, T. **DAMERAU**, F. **WEISS**,"TEXT MINING PREDICTIVE METHODS FOR ANALYZING UNSTRUCTURED INFORMATION,**" SPRINGER**, 2005.
[2] L. Talavera and J. Bejar,"Generality-Based Conceptual Clustering with probabilistic concepts," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no.2, pp. 196-206, Feb. 2001.
[3] Sergio Bolasco ,AlessioCanzonetti , Francesca Della Ratta-Rinald and Bhupesh K. Singh, (2002), "Understanding Text Mining:a Pragmatic Approach", Roam, Italy.
[4] R. Swan and J. Allan, "Automatic Generation of Overview Timelines", In SIGIR2000, 49-56.
[5] Andreas Hotho, Steffen Staab, GerdStumme,"Text Clustering Based on Background Knowledge", Technical Report 425, Inst. AIFB, Univ. of Karlsruhe, Apr. 2003.
[6] M. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, no. 3, 1980.
[7] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.
[8] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern-Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.
[9] http://en.wikipedia.org/wiki/Tf*idf
[10] K.K. Bun and M. Ishizuka, "Topic Extraction from News Archive Using TF_PDF Algorithm," Proc. Third Int'l Conf. Web Information Systems Eng. (WISE '02), pp. 73-82, 2002.
[11] X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '03), pp. 587-594, 2003.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

[12] S.T. Dumais, "Improving the Retrieval of Information from External Sources," Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.
[13] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report Raport NR 941, Norwegian Computing Center, 1999.
[14] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization," Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997.
[15] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management: An Int'l J., vol. 24, no. 5, pp. 513-523, 1988.
[16] Kuan-Yu Chen, LuesakLuesukprasert, and Seng-cho T. Chou," Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 8, pp. 1016-1025, AUGUST 2007.
[17] Y. Yang, T. Pierce, and J. Carbonell., "A Study of Retrospective and Online Event Detection. In Proceedings of the 21st Annual International ACM SIGIR Conference," Melbourne, Australia. ACM Press. 1998, 28- 36.
[18] B. Thorsten, C. Francine, and F. Ayman.," A System for New Event Detection.," In Proceedings of the 26th Annual International ACM SIGIR Conference, New York, NY, USA. ACM Press. 2003, 330-337.
[19] K. Zhang, J. Li, and G. Wu., "New Event Detection Based on Indexingtree and Named Entity," In Proceedings of the 30th Annual International ACM SIGIR Conference, Amsterdam, the Netherlands. ACM Press. 2007, 215-222.
[20] J. Allan, R. Papka, and V. Lavrenko.,"On-line new event detection and tracking," In Proceedings of the 21st Annual International ACM SIGIR Conference, Melbourne, Australia. ACM Press. 1998, 37-45.
[21] C. Wang, M. Zhang, S. Ma and L. Ru.,"Automatic online news issue construction in Web environment.," In proceedings of the 17th international conference on World Wide Web, 2008, 457-466.
[22] Yan Gao Jin Liu and PeiXunMa ,"The Hot KeyPhrase Extraction based on TF*PDF," IEEE, 2011.
[23] M. Lan, C.L.Tan and Hwee-Boon Low, "Proposing a New Term Weighting Scheme for Text Categorization," American Association for Artificial Intelligence, pp. 763-768, 2006.
[24] Joel, Thomos, Brain, Mark and Ali,, "TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams."
[25] Tian and Tong, "An improvement to TF: Term Distribution based Term Weight Algorithm", IEEE Computer Society, pp:252-255, 2010.
[26] Stephen Marsland, Chapman & Hall, Machine learning An Algorithm Perspective, CRC. 2009.
[27] Shady Shehata, FakhriKarray, and Mohamed S. Kamel," An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, pp.1360-1371, OCTOBER 2010.
[28] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.
[29] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern-Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.
[30] Evan Wei Xiang, Bin Cao, Derek Hao Hu, and Qiang Yang," Bridging Domains Using World Wide Knowledge for Transfer Learning", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 6, pp. 770-783, JUNE 2010
[31] Paul Kingsbury and Martha Palmer," PropBank: the Next Level of TreeBank", Proc. Workshop Treebanks and lexical Theories,2003.
[32] Kowalski, Gerald, Mark T Maybury, Information Storage and Retrieval Systems: Theory and Implementation, Springer.
[33] Ricardo Baeza-Yates, Modern Information Retrival, Pearson Education, 2007.
[34] David A Grossman and OphirFrieder, Information Retrieval: Algorithms and Heuristics, Second Edition, Springer.
[35] William B Frakes, Ricardo Baeza-Yates, Information Retrieval Data Structures and Algorithms, Pearson Education, 1992.
[36] Robert Korfhage, John Wiley & Sons, Information Storage and Retrieval.
[37] Ethem. L. Padin, Machine Learning,
[38] Stephen Marsland, chapman and Hall/CRC, Machine Learning An Algorithm Perspective, 2009
[39] R.O. Duda, P. Hart, David G. Stork, Siley, Pattern Classification 2nd edition.
[40] Smith Tsanf, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau Dan Lee, "Decision Trees for Uncertain Data", IEEE Trans. Knowl. Data Eng., vol. 23, no. 1,pp. 1-15,2010.
[41] A Survey on Transfer Learnig, *www1.i2r.astar.edu.sg/~jspan/publications/TLsurvey_0822.pdf*
[42] G.R. Xue. W.Dai. Q.Yang, and Y.Yu," Topic-Bridged PLSA for Cross-Domain Text Classification", Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08),pp. 627-634, July 2008.
[43] J. Jiang and C. Zhai, "Instance Weighting for Domain Adaptation in NLP," Proc. 45th Ann. Meeting o the Assoc. for Computational Linguistics, June 2007.
[44] A. Argyriou, C.A. Micchelli, M. Pontil, and Y. Ying, "A Spectral Regularization Framework for Multi-Task Structure Learning", Proc. 21st Ann. Conf. Neural Information
[45] Domeniconi, and J. Hu, "Cross-Domain Text Classification Using Wikipedia", The IEEE Intelligent Informatics Bull., vol. 9, pp. 5-7, Nov. 2008.
[46] WordNet, http://www.cogsci.princeton.edu /~wn/, 2006.
[47] T. R. Gruber," A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition," 199-220, 1993.
[48] R. Studer, V. R. Benjamins. FenselD.,"Knowledge Engineering, Principles and Methods. Data and Knowledge Engineering," 161-197, 1998.
[49] N. Guarino. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration [M].,Pazienza M. T., Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Berlin/Heidelberg: Springer, 1997: 139-170.
[50] Feng Hu, Yu-fengZhang, Text Mining Based on Domain Ontology, IEEE, 1456-1459, 2010.
[51] S. Bloehdorn and P. Cimiano and A. Hotho and S.Staab, "An Ontology-based Framework for Text Mining," 1-20, july 2004.
[52] L. P. Jing, L. X. Zhou, et al.," Ontology-based Distance Measure for Text Clustering "2006.
[53] Automatic Summarization. Wikipedia, http://en.wikipedia.org/wiki/Automatic_summarization.
[54] ShanmugasundaramHariharan," Extraction Based Multi Document Summarization using Single Document Summary Cluster", Int. J. Advance. Soft Comput. Appl., Vol. 2, No. 1, pp. 1-16, March 2010.

Y. Jahnavi: received her B.Tech (Computer Science &Engineering) degreefrom JNTU and M.Tech (Computer Science and Technology) degree from SVUniversity. She is currently an Associate Professor in the Department of Computer Science &Engineering at GIST. She is working for herPhD in the area of Text Mining. Her current research interest includes Data Mining, Information Retrieval, Machine Learning, etc.

Dr. Y. Radhika: received her B.E (Civil Environmental Engineering) ,M.Tech (Computer Science and Technology)andPhD degrees from AndhraUniversity. She is currently an Associate Professor in the Department of Computer Science& Engineering at GITAM University.She has been publishing in national and international journals. Her current research interest includes Data Mining, Neural