ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

An Extractive Summarization of Document Using Conceptual Mining and Sentence Ranking

Pragya Lodhi[1] , Prof. Tripti Sharma[2]

M.Tech Student, Dept. of CSE, RCET, Bhilai, Chhattisgarh, India[1]

Associate Professor, Dept. of CSE, RCET, Bhilai, Chhattisgarh, India[2]

ABSTRACT: In this digital era all most every information is presented in and preferred to store in digital form. It is laborious and difficult task for a reader to go through the whole large documents quickly. This problem can be solved by getting an outline or the summarization of the large document. Like, before starting a project or a research work the students or a researcher have to read many research papers. Some time it may happens that the abstract and conclusion don't contain the actual content through which the reader could come to know that what is written in the whole paper, so a summarization of the paper can ease the work of the reader. Summarization can be of two type extractive and abstractive. In this paper the model has been projected which uses conceptual mining, ontological matching, pattern mining and ranking of sentences for extractive summarization. An enhanced text clustering approach can be achieved by Conceptual Rule Mining in document which detects the influential and related term and phrases that influence the topic of the document. The procedure of vector space model is divided into various steps. The first step is the indexing of the given document which results in the extraction of content bearing terms present in the text document. Ontological matching is used to detect the synonym, hypernym and hyponym of the word. POS tagger is used to detect verb, noun and adjective part of speech present in the text document. The next step is the assignment of weight to the indexed terms to improve the retrieval of document which are relevant to the user after which pattern mining is performed to detect the significant pattern in the document. The last step is hybrid ranking of the sentences with respect to the weight assigned to the sentences.

KEYWORDS: Concept mining, ontological matching, extractive summarization, Ranking, pattern mining**.**

I. INTRODUCTION

The Natural language processing is a discipline of linguistics, artificial intelligence, and computer science which deals with the communications among the human and computers. Particularly it deals with the extraction of useful information from the data using natural language processing and other techniques. Natural Language Processing can be defined as a method of probing and evaluating the states of words in document. Similarly Text mining attempts to find out the contemporary and formerly indistinct information with help of methods present in data mining. There is a contemporary technique in which concept mining is employed with natural language processing in addition to geometric analysis for text mining. Impending natural language processing, to avert the vagueness of various meaning of a different terms and number of depiction for a unlike sense NLP may be worn. In proposed model a novel synonym based concept mining model is projected. It supersede to entire remuneration of extant concept based mining model.
To find out the resemblance among the documents a concept-based similarity measure is used. Concepts get across through the local context information, to find out an exact resemblance among the text documents. The concept based comparison measure, which depends on identical concepts in the sentence, document, corpus and collective approach instead of on the single and independent terms solely, has been constructed. The concept-based resemblance measure depends on mainly three significant features. Firstly, the examined tagged terms are concepts that confine the semantic structure of every sentence. Secondly, to compute the involvement of the concept in the meaning of the sentence, likewise to the subject of the document the frequency of a concept is used. Lastly, the numerous documents which contain the examined concepts is used to distinguish between documents in computing the likeness.

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Vector Space Model can be defined as an algebraic model for presenting the text documents as vectors of identifiers, like, index terms. VSM is employed for information filtering, indexing retrieval, and relevancy rankings. The process of VSM can be described in following steps. Firstly all the useful and significant words that are present in the document are extracted. Then for next step weight is assigned to the extracted terms. Lastly, ranking of the documents or sentences is done on the basis of weight assigned to it.

Concept Mining is a technique that is employed to retrieve the concepts which are present in the document. The concept could be words, phrases and are entirely reliant on the semantic structure of that sentence. The resemblance measure is used for concept analysis is may be employed on the different levels like sentence level, document level or on corpus level. The concepts are initially retrieved with the help of the part of speech tagger and examined with respect to the sentence, document, and corpus levels. Whenever the concepts got corresponding in not related documents, it generates errors in form of noise. Thus, while document resemblance is computed, the concepts turn out to be insensitive to noise

The massive quantity of information available nowadays has lead to information burden problem. For a reader it becomes very difficult to go through the large quickly. So document summarization is one of the feasible solution to get the rid of this information surplus trouble. Summarization is a hard problem of Natural Language Processing because, for faultless performance, it is essential to get the point of a text. This demands semantic examination, dissertation processing, and inferential explanation. Document summarization is the process of extracting out content from the text document, and get the most meaningful and useful content to the user in a precise form. This process reduces the problem data surplus as just a quick read is required to understand the document in place of analysing the whole document. Summarization process can be of two type extractive summarization and abstractive summarization [10].

II. RELATED WORK

Text Mining is considered as a very significant topic among researchers. Text Mining is the detection of novel, formerly unidentified information, automatic extraction of information from different resources. [4] This paper contains, Text Mining techniques survey and text mining applications of has been presented.

To improve the text clustering technique, Conceptual Rule Mining technique is used to assess the connected and significant sentences that are contributing to topic of the document. The concept-based mining model has capability to efficiently determine between significant and non significant terms with respect to word or phrase which embrace the concepts that symbolize the sentence meaning and the sentence semantics structure [2]. Weight is assigned to the sentences through conditional probability. To diagnose the sentence resemblance from which unique sentence connotation contributing to the document topic is listed, Probability ratio is used. After this sentences ranking is implemented that is computed by employing the weights allocated to each and every sentences [1]. Various application of concept based mining model for text clustering has also been described [6].

[3] A single document summarization technique has been projected in the paper which employs the two sentence importance measures: occurrence of the every noteworthy term within the sentence and resemblance to the other sentences. The ranking of the sentences are done on the basis of their weight which are assigned to it and the top ranks sentences are selected for summary. The summary is assess by using 'recall' evaluation measure. [4]This paper made a apparent and easy outline of functioning of text extraction from Pdf document in discrete process. Numerous text extraction systems are present now a days but researchers are researching in this subject to get better efficiency because still we find trouble in retrieving the text from the document which contains images, tables, diagrams and other things. So the main idea is to retrieve the text present in composite document efficiently and effortlessly.

[5]Till date a huge number of ranking algorithms have been projected. This paper has summarized and compared some of these algorithms. even though every algorithm has its pros and cons, but each algorithm is evaluated using various evaluation measures to find the best algorithm.

Text mining is an imperative technique in the field of data mining which incorporate very flourishing method to retrieve the valuable patterns.[9] The paper targets on budding an effective technique for dig out the significant patterns present in the document. Pattern mining approach is employed to discover text patterns, such as frequent item sets, co-occurring terms, closed frequent item sets. [7]In spite of generally employed keyword-based approach, the pattern based model which consists of frequent sequential patterns is used to achieve the similar concept of tasks. In this anticipated method two techniques which rely on the use of pattern deploying strategies is used. The methods

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

implement the mining sequential pattern procedure that discover semantic patterns from the text and arrange these patterns using projected deploying algorithms. [8]This projected model, an efficient pattern detection method has anticipated to remove the misinterpretation and low-frequency problems which generally occur in text mining. Projected method uses two techniques, pattern evolving and pattern deploying, to process the exposed patterns in text documents. The projected model performs so well in other the concept based model and pure data mining-based process, and also in term based state of art models.

III. PROPOSED METHOD

Concept based mining approach:
In this projected concept based model a test text document of any four topic (computer network, data mining, machine learning, image processing) is given as input, and the output will be the topic detection of the document to which it belongs between the above stated topics and an extractive summary of the document. The given text document has distinct sentence boundaries. Every sentence present in the document is labelled and processed individually. Text mining can be defined as a procedure of converting the text into numeric form. Which means, every term or word that dwells in the test document will get indexed and counted for computing a table of documents and words, i.e., to specify the number of times that every term occur in the document a matrix of frequencies will get created. This essential course of action can again refined to eliminate some of the ordinary terms like "a" and "the" which are generally known as stop word and to convert the words that are in different grammatical forms into its root word such as converting "processing," "processed," to "process" etc. This process is known as stemming. When a table of unique words or terms in documents is derived, all data mining techniques and customary statistical can be practice to obtain dimensions or clusters of words or documents, or to detect the "significant" words that efficiently foresees other outcome variable of interest. After the process of stop word and stemming, each remaining word will automatically get tagged with the part of speech through POS tagger. In this tagging only the noun, verb and adjective arguments get labelled and considered as a concept.

Concept Analysis Sentence-Based:
A novel concept-based frequency measure called conceptual term frequency (ctf) is anticipated to examine every concept at the sentence level. The ctf computation of concept c in sentence s, and document d are as follows:
Ctf computation in Sentence s: The ctf can be defined as the total existence of concept c in argument structure sentence s. The concept c, which repeatedly appears in other argument structures of that sentence s, has the primary role of adding meaning to s. In this scenario, the ctf is a confined measure on the sentence level.
ctf computation in Document d : It may happen that a concept c may have various ctf values in other sentences in the same document d. Thus, the ctf value of concept c in
document d is calculated by:

$$ctf = \sum_{n=1}^{s_n} ctf_n / s_n$$

Here s_n indicates the number of sentences that include concept c in document d. To determine the involvement of concept c to the significance of the sentences ,the average calculation of the ctf values of concept c in its sentences of document d is measured. A concept which has ctf values in the majority of the sentences in a document, have a foremost involvement in the meaning of the sentences and which get to determine the subject of that particular document. Thus, computing the average of the ctf values determines the overall significance of every concept to the semantics of a document through the sentences. The computation of ctf in a document is illustrated; suppose a concept c that occur thrice in document in the fourth and the fifth lines. The concept c comes five times in the argument structures of the fourth sentence , and four times in the argument structures of the fifth sentence. In this case, the ctf value is calculated as follow:

(4+5+3)/3=4

Concept Analysis Document-Based
For analysing every concept at the document level, the computation of *tf* which is concept based term frequency is done. The *tf* is computed on the basis of the occurrences of concept *c* in the test document *d*. The *tf* is a measured at document level.

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Concept Analysis Corpus-Based

To mine the concepts which could distinguish between documents, , the number of documents hold the concept c, the concept-based document frequency *df* is computed. The *df* is a comprehensive measure on the corpus level. As these concepts can discriminate the documents among others, *df* is employed to compensate the concepts which occur in a very few documents.

The computing the ctf, tf , and df measures in a corpus is obtained by the following algorithm :

Step 1: d_{doci} is a test document
Step 2: L is empty matched concept list
Step 3: s_{doci} is a new sentence in d_{doci}
Step 4: Build concept list C_{doci} from s_{doci}
Step 5: **for** each concept $c_i \in C_i$ **do**
Step 6: Compute ctf_i of c_i in d_{doci}
Step 7: Compute tf_i of c_i in d_{doci}
Step 8: Compute df_i of c_i in d_{doci}
Step 9: d_k is training document, where k= {0,1,…,$doci$ -1}
Step 10: s_k is a sentence in d_k
Step 11: Build concept list C_k from s_k
Step 12: **for** each concept $c_j \in C_k$ **do**
Step 13: **if** $(c_i == c_j)$ **then**
Step 14: update df_i of c_i
Step 15: compute $ctf\ weight = avg(ctf_i, ctf_j)$
Step 16: add new concept matches to L.
Step 17: **end if**
Step 18: **end for**
Step 19: **end for**

The following is the flow chart of proposed methodology

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

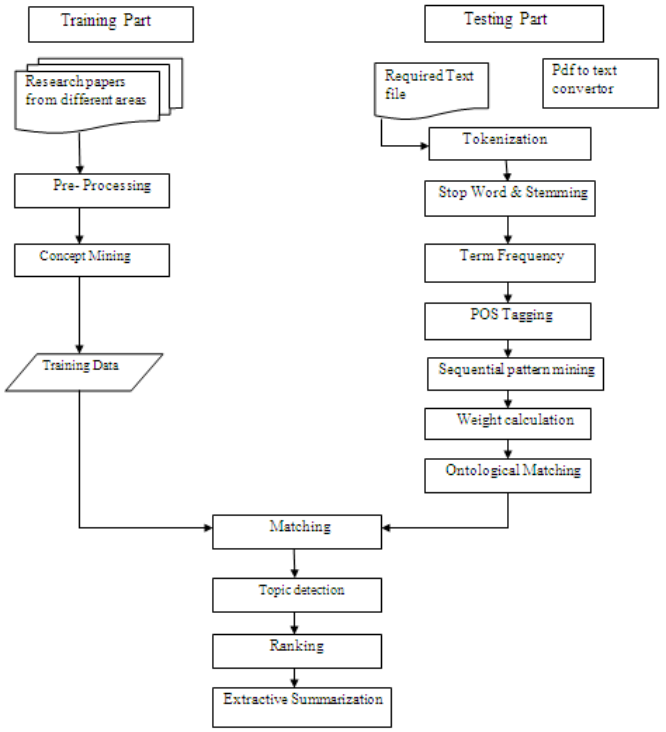


Fig: Proposed Methodology Flowchart

Closed Pattern paragraph: Pattern mining is used for efficient extraction of significant patterns which can be pattern, term or phrase. It is a complicated work to construct a procedure that uncovered the patterns in test documents for information filtering systems. Suppose given a term set X in document d, dp denote the paragraph set such that $dp \in PS(d)$. Such that $X \subseteq dp$, i.e. $X = \{dp | dp \in PS(d), X \subseteq dp\}$.

The paragraph set dp with respect to term set X is called frequent pattern if its support $sup \geq min\_sup$. Algorithm for pattern analysis is as follows.

Input: positive document D; minimum support min_sup
Output: pattern paragraph with respect to term DP and support of term
Step 1: $DP = \emptyset$;
Step 2: **For each** document d $\in D$ **do**
Step 3: Let PS(d) be the set of paragraph in d;
Step 4: $SP = PMining(PS(d), min\_sup)$;
Step 5: **End For**

Topic detection of the document is performed with the help of dictionary of data mining, image processing, machine learning and computer network. The extracted terms or concept is matched with the words in dictionary to detect that particular term belongs to which topic. For example protocol, packets, IP address words belong to the computer network.

So the words get tagged with the topic to which they belong. At last the topic is decided by summing all the four tags individually and the greater will become the topic of that particular document.

Ranking of the sentences

To improve the efficiency of the ranking function of sentences a hybrid approach can be used. K- mean clustering is used to cluster the sentences on basis of highly contributive sentences, medium contributive and low contributive sentences. One of the ranking approach is as follows:

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Input: Document D, Set of sentences S, Ranking function, cluster number K
Step 1: t – 0
Step 2: Get the initial Partition of s i.e. C_k^t; k= 1, 2… k, calculate cluster center expressed as,

$$Center\ C_k = \sum_{S \in Ck} S_i / |C_k|$$

Where$|C_k|$ is the size of C_k
Step 3: **For** (t=1; t<number of items &&ε >0.001; t++)
Step 4: Calculate the cluster ranking
Step 5: Get new attribute for each sentence S,
Step 6: **For** each sentence s_i in S
Step 7: **For** K=1 to k
Step 8: Calculate similarity value $sim(s_i, C_k^t)$

$$sim(s_i, C_k^t) = \frac{\sum_{i=1}^{K} s_i(i) center C_k(i)}{\sqrt{\sum_{i=1}^{K} s_i(i)^2} \sqrt{\sum_{i=1}^{K} Center C_k(i)^2}}$$

Step 9: **End For**
Step 10: For each sentence s_i in S
Step 11: **For each** K=1 to k
Step 12: $f(s_i) = \sum_{k=1}^{k} \alpha_k . r(s_j | C_k)$
Step 13: **End For**
Step 14: **End For**

The output of this algorithm will be the sentence final ensemble ranking vector *f* (*S*). Another approach is simply by adding the entire concept present in all the sentences. The sentence with highest number of concept in it will be considered as the highest ranked sentence.

IV SIMULATION RESULTS

This section presents the results for the evaluation of the proposed approach. Through this projected model it has been shown that how significant sentences and terms which contribute to topic of document are detected by employing conceptual rule mining followed by pattern mining and ranking. It illustrate the procedure of the retrieving the sentences which are more relevant to the topic to make an extractive summary which help the user to get a quick review of the document.
The performance of the projected model i.e. extractive summarization of the document using Conceptual mining and sentence ranking is measured in terms of :

1. Accuracy
2. Precision rate
3. Recall

The formula used for Accuracy , precision and recall are as follows:

$$Accuracy\ \frac{True\ Positive + True\ Negetive}{Ture\ Positive + True\ Negative + False\ positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

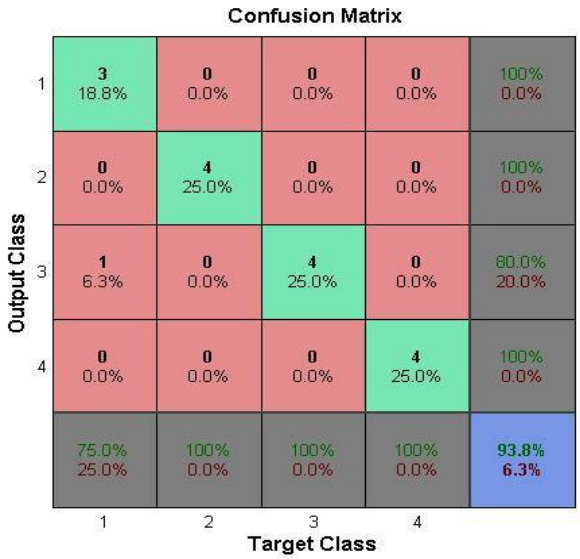
The performance of projected model is shown with the help of confusion matrix.

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016



This confusion matrix output represents that the topic identification of test document for 3 among 4 classes is always accurate, but for the fourth class it may detect the topic incorrectly. In confusion matrix the accuracy is shown which is 93.8%, and the previous method has 89% accuracy rate. The performance of topic detection can be increase by improving or adding the more terms to dictionary.

Precision & accuracy of the proposed model is as follows:

No. of doc	Precision	Recall
4	1	1
8	1	1
16	0.882	0.937
32	0.852	0.906
40	0.790	0.85

The above table show that as the number of document increases, the value of recall and precision decreases.
Similarly the average precision rate of the of the proposed document is 91% and the average precision of existing document is 88%.

V CONCLUSION AND FUTURE WORK

Projected model overpass the space among natural language processing and text mining authorities. By using the semantic construction of the sentences in documents, an improved text extraction outcome and effective summarization can be obtained. A novel concept based mining model composed of components which are as follows. Firstly the concept is analysed at sentences level using the ctf measure, secondly the concept analysis is done in the document level i.e. df which shows the number of concept present in the document. Thirdly the concept is analyses at corpus level. Pattern mining technique is used to deploy the useful patterns present in the document which improve the term selection process. At last ranking of sentences is performed. To improve the ranking output a hybrid approach is used. To extend this paper one can employ the same model to text classification. The purpose is to examine the practice of such model for extractive summarization and effect in concept mining and ranking performance, in contrast to that of conventional methods.

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

REFERENCES

[1] V. M. Navaneethakumar, "Mining Conceptual Rules for Web Document Using Sentence Ranking Conditional Probability" International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME) February 21-22 2013 IEEE
[2] M. Yasodha, Dr .P. Ponmuthuramalingam, "An Advanced Concept-Based Mining Model to Enrich Text Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012
[3] Y. Surendranadha Reddy, Dr. A.P. Siva Kumar, "An Efficient Approach for Web document summarization by Sentence Ranking." 2012, IJARCSSE
[4] Deepak Motwani , A.S. Saxena, "Multiple Document Summarization Using Text-Based Keyword Extraction" Springer Science+Business Media Singapore 2016
[5] Shashank Gugnani, Tushar Bihany, Rajendra Kumar Roul, "A Complete Survey on Web Document Ranking," International Journal of Computer Applications Volume ICACEA - No. 2, 2014
[6] Modu Sowjanya, K.Ravindra, Y.Ramesh Kumar,"Application of Concept-Based Mining Model in Text Clustering", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014
[7] Sheng-Tang, Wu Yuefeng, Li Yue Xu, "Deploying Approaches for Pattern Refinement in Text Mining", the Sixth International Conference on Data Mining (ICDM'06) IEEE 2006
[8] Ning Zhong, Yuefeng Li, and Sheng-Tang WuI, "Effective Pattern Discovery for Text Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012
[9] V.Aswini, S.K.Lavanya," Pattern Discovery for Text Mining", 2014 INTERNATIONAL CONFERENCE ON COMPUTATION OF POWER, ENERGY, INFORMATION AND COMMUNICATION (ICCPEIC)
[10] Vishal GuptaS, Gurpreet Singh Lehal," A Survey of Text Summarization Extractive Techniques.", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010

BIOGRAPHY

Pragya lodhi is a M.Tech student of the Computer Science and Engineering Department, in RCET Bhilai. She received Bachelor of engg.(B.E.) degree in 2014 from S.S.I.T.M, Bhilai, C.G., India. Her research interests Data mining and Big data.

Tripti Sharma is a Associate professor of the Computer Science and Engineering Department, in RCET Bhilai. She completed her B.E and MTech in Computer Science and Engineering branch. Her research interests Data mining and Image processing.