A Survey on Mining Conceptual Rule and Ontological Matching for Text Summarization

Pragya Lodhi[1] Tripti Sharma[2]
[1]M.Tech. Student [2]Associate Professor
[1,2]Department of Computer Science & Engineering
[1,2]Rungta College of Engineering & Technology, Bhilai

Abstract— The escalation of web information has constrained rigorous research in the area of text summarization in Natural Language Processing community. When an information is being retrieved from such an enormous collection of web documents, thousands and lakhs of documents are retrieved daily. Hence, for user, it's impossible to read all the retrieved documents. So for abetting and elucidate text information the automatic text summarization technique has become a very supreme and felicitous tool. The technique of Summarization consist of curtailing a text document with a computer program to create a outline that retains the important details and overall meaning of the original document. Today text mining has emerged as a very popular innovative field among researchers that endeavors to extract important and useful information from natural language processing of text. Various mining model are present that can be use that identify the concepts of the given document, phrases or sentences which identify the topic and summary of the given document.

Key words: Conceptual Rule Mining, Text Clustering, Conditional Probability, Concept-Based Mining Model, Ranking, Extractive Summary, Ontology Matching, Semantic Similarity, Natural Language Processing

I. INTRODUCTION

The Text Mining technique usually used to process unstructured data, dig out useful numerical facts present with in the document, and provide the useful and important data enclosed in text attainable to the different algorithms present for data mining. Information is extracted to get the outline of the text contained in the document or to summarize the document on the basis of the terms or words comprise in it. It allow you to analyze the terms, clusters of words within the document, we could also analyze documents to detect the similarities or decide how much they are related to each other considering the words that are useful.

Statistical analysis of a word and terms is a method on which some of most common techniques that are used in text mining are based. To know the importance of a particular term within a document Statistical analysis of a term frequency can be used. Sometimes it may occure that two or more terms with in the document can have a same frequency, but in that specific one term contributes more to the sementic of the sentences as compared to alternative terms. Generally while considering text mining techniques the basic parameter such as term frequency is basis for computing the significance of the word in a particular document. However the problem with statistical analysis is that, the actual meaning of word may not have the accurate meaning of that word. So a new technique which depends on conceptual mining rule model and approach based on synonym is used to efficiently search similar words and associated topic of the documents can overcome the above stated problem. The relations between part of speech like verb, adjectives or noun and the related arguments in the sentence can examine the terms that exist in a sentences of the document. Data regarding to what is happening to whom or by whom, illustrate the effectiveness of each and every word or term within a sentence and analyze the theme of each sentence of the document and determine which sentence contribute more to the topic. These work highly contributes to natural language processing (NLP) and text mining research field. An efficent and effective text cluster can be obtained by considering and exploring the linguistics structure of sentences.

The ontology matching generally deals with the vocabulary or synonym that describes a domain of interest and a specification of the meaning of terms employed in the vocabulary. Depending on the exactness of the definition, the conviction of ontology encompasses various conceptual models, for classifications, clustering, fully axiomatized theories and database schemas.

The concept based mining that is used to examines the terms within the phrases or sentence in the collection of document is applied. The concept based mining model effeciently diffrentiate between unimportant words and important words in reference to linguistics of the sentences and those words that has the concepts which contains the actual meaning of sentence. The model that has been projected consist of concept-based similarity measures, concept analysis based on sentence, concept analysis based on document and concept analysis at corpus level. Various experiments have demonstrate the noticable improvement in the quality of clustring using sentence based, document based, corpus based and combined approach concept analysis. To get the similarity between documents a new similarity measure has been proposed that can also be used for classification and clustering of the documents with in existing clusters.

Text summarization is a way to abridge the abundent information into a brief form by selecting the imporatant and related data and discarding unimportant and redundant data. Generally there are two approaches of text summarization, that are: extractive summarization and abstractive summarization. The abstractive summarization contains the understanding of text in the document by linguistic technique to interpret and examine the text. The abstractive summarization aims to provides a generalized summary, transference in information in a very concise manner, and typically needs language generation and compression techniques. However, the extractive summarization extract the noticeable phrases or sentences from the source documents and cluster them to provide a summary without changing the source text. Usually,

sentences are in the same order as within the original document.

II. RELATED WORK

Today various advanced applications has been created for text mining, like article summarization, essay summarization, queries responsive, and in shorts: news summary.[15] It has been observed that statistical text-mining techniques are not that effective for these kind of several applications , since the morphological structure of the data is not used. Thus, natural language processing techniques approach is used, in which firstly the given text is parsed and patterns are used for mining purpose and analysis of the trees is performed which is generally a very complicated task. In semantic similarity method the similar words or terms that exist in the word-pairs are analyzed. The degree of relationship that exist between the word-pairs can detect by using these measures.[3] In projected methodology the similarity among the words detected on the basis of page counts and snippets based lexical pattern clustering. Clustering and classification augmented with semantic similarity (CCASS) is introduced For calculating the linguistic similarity between a word-pair,. CCASS is one of the new techniques that use text snippets and page count returned by search engine. Various measures to detect similarity has been elucidate by considering the lexical pattern clustering from snippet and page counts of word pairs. The Lexical pattern clustering is applied on search engine obtained text snippets which are given to the support vector machine (SVM). SVM evaluates the linguistic similarity among word-pairs. After this K-Means clustering can be used for creating clusters according to the value we get from support vector machine. Forthcoming word-pairs will be classified, after evaluation of their linguistics similarity. A new cluster will be created if it does not match with the prevalent clusters.

[5]The mentioned paper describes the planning of a system for extracting key sentences or phrases from a unique document. The principle of the algorithmic rule is to cluster the phrases or sentences of the documents to focus on elements of data that are semantically equivalent. The clusters of phrases or sentences, that engraves the main subject of the document, are then analyzed to find out the main topics of the document. At last, the most useful keywords, or groups of words, from these topics are featured as key phrases.

Key phrases can be seen as words, or groups of words, that frames the key concepts of a document. Key phrase represent useful data regarding a document and represent an alternative, or a complement, to full-text assortment. Pertinent key phrases are useful to potential readers who can have a fast overview of the content of a document and might choose easily which document to read. Presently, the leading powerful key phrases extraction algorithms are based on supervised learning. These strategies focus on the matter of associating

Key phrases to documents as a classification process. However, the actual fact that this approach needs a corpus of comparable documents, that isn't continually readily offered, constitutes a significant disadvantage. For example, if one reads a new net document, one would possibly wish to understand most topics addressed in that document as quickly as possible. So in this case domain-independent keyword extraction technique is required which is applied to a single document. [9] To increase the coherence of the extracted key phrases, improvements should be done to the key phrase extraction algorithm. This will includes the usage of the degree of statistical association among candidate key phrases to affirm that they will be semantically connected. Web mining is used to measure the statistical association.

In a contemporary mining model based on concept, four basic component is introduced, that are suggest to increase the quality of text clustering has been discussed.[6] By utilizing the linguistics structure of the sentences within documents, an more efficient and effective text clustering result can be attain. The concept analysis based on sentence is the primary element that is used to examine the semantic structure of every sentence to get the concepts of the sentences by considering the measure based on conceptual term frequency (ctf). The concept analysis based on document level is the second component that is used to examine each and every concept using the concept-based term frequency, tf at document level. The analysis of concepts on the corpus level is the third component which is done by using the document frequency df measure. The last component deals with similarity measure based on concept that measures significance of every concept in reference to the linguist of sentence, the subject of the document, and the differentiate between the documents collection. While considering all the parameters that affects the concept's weight in sentence or at corpus levels, a similarity measure which is based on concepts is capable of correct calculation of pair wise documents is formulated.[7] The concept – based similarity function implementation is used where document equivalence is used to cluster the documents as an example in applications that cluster newspaper articles for topic detection and chase. The concept primarily based similarity function is used for web document clustering.

To detect the related concepts that raised in the sentences in the document, a model is intended based on appropriate weights computation by using the conditional probability. [1] To check the similarity between the sentences that are contributive to the topic probability ratio is used and those sentences which contribute more to the topic are selected. For text mining the conceptual rule worked under two totally different phases. At first part it describes the generation of conceptual rules for the sentence meaning and sentences that are connected. The weights that have most contribution to the subject of the document are used for additional method in concept based mining model. Second part describes the analysis of conditional probability for sentence weights. Then for checking the similarity between the sentences the probability ratio is calculated that distinctive the sentence which contributes more to the theme of the document are selected.

Mining conceptual rules from a web documents is a 3-step process which consist of 1) pre-processing, 2) the actual mining and 3) post-processing. At first preprocessing step the web documents are regenerated from exterior documents into widespread illustration. The important and useful words, concepts or keywords are mined, and after that different functions could be done in text for improve the standard of output or reduce the execution period of

analyzing the mining method. After that the real process of mining is conducted, which produces the several conceptual rules. Quantity of rules could be higher, which rules are most fascinating supported by some measure and which rule should be discarded is decided in the post processing phase.

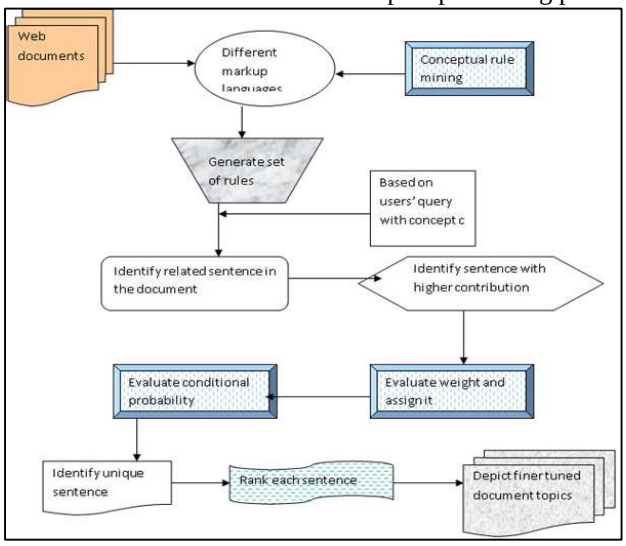

Fig. 1: Architecture Diagram of the CRMSRCP

Concept hierarchy is the key concept of ontology,[8] and also the concept hierarchy acquisition is been a subject undergoing intense study within the area of ontology learning. The proposed paper projects a hyponymy identification technique of domain ontology concept based on cascaded conditional random field(CCRFs) and hierarchy clustering. It accepts the text as extracting object, applies CCRFs to identify the domain ideas. 1st the low layer of CCRFs is used for spot easy domain concept, then the outputs are sent to the upper layer, where the nesting concepts are recognized. After that we tend to adopt hierarchy cluster to spot the hyponymy relation between domain ontology concepts. The experimental results demonstrate the planned technique is efficient.

For efficient clustering [13] ranking of sentences has so far been considered as a severe problem. On account documents typically cowl variety of themes, with every topic or a theme delineate by a cluster of extremely similar or connected sentences are clustered. Clustering of sentences was recently probe so as to generate a more accurate summary. The presented ranking technique which is a cluster based approaches applies both ranking and clustering separately. Which leads to the ranking performance are going to be inescapably influenced by the clustering result.

For ranking there are various algorithms like[14]:

- Page Rank Algorithm : It Calculates the values at index time and the priority based pages or sentences are represented as a result
- Weighted Page Rank Algorithm: Evaluate the values at index time and the results are sorted on the basis of Page or sentence relevance.
- HITS Algorithm: 'n' number of remarkably relevant pages or sentence is computed and finds values on the fly.
- Distance Rank Algorithm: Compute the Minimum Average Distance Between two and more pages or sentences
- Eigen Rumor: With the help of agent to object link the adjacency matrix is constructed for use.

Ranking can be done through various techniques[1] ranking according to rank-score, ranking as per path-rank,. Since only the rank-score technique might not be able to distinguish between different types of Clusters, path-rank Technique can be used. we tend to outline the path-rank because the distance between the desired cluster and the matched term according to the hierarchy of classification ontology.

On the basis of semantic similarity value obtained, the clustering of documents are performed by considering their probability values. A K Means clustering algorithm or Vector Space model can be used to form clusters. The simple K Means is a type of segregation clustering technique. K Means clustering can be considered as an iterative algorithm where, until the required set of information is assigned to the appropriate sets of clusters. Whereas for document clustering method and symbolizes knowledge for content categorization and clustering the Vector space Model (VSM) is widely used. The feature vector that is symbolized in the document could be a phrase or a word. Every feature vector is consigning the weight supported on the term prevalence of the terms contained in doc. To get the correspondence among the documents, similarity procedures which accepts feature vector are used.

Extractive summarization picks out very appropriate sentences in whole document and too preserves the minimum repetition in summaries. Some of the major techniques for extractive summarization are Term Frequency-Inverse Document Frequency (TFIDF), Graph theoretic approach, text summarization based on concepts, regression based Text summarization, Query based extractive text summarization.

III. CHALLENGES

Some of the most serious problem in the area of text mining are due to the intricacy of a natural language. In natural language processing ambiguity is a most challenging problem. Ambiguity can be referred as the doubtfulness or uncertainty as regarding to interpretation. In a text document one word can have more than one meanings and one phrase or sentence are often understood in several different ways which may led to numerous meanings of statement. Even though several researches has been performed to resolve the ambiguity drawback but the problem still don't have any perfect solution. There are some problems in extractive summarization also which are:

1) Extracted sentences typically becomes longer than average. So the segments that don't seem to be useful in the summary are also get inserted.
2) Useful and significant data are mostly spread over all the document, and cannot be captured by an extractive summaries
3) Conflicting data might not be represented meticulously.
4) Fine extraction usually results in issues in overall coherence of the outline

IV. CONCLUSION

This review paper discusses the related work that has been done in the field of text mining, Mining of conceptual rules, Ontological Matching, Ranking and Extractive summarization. Number of strategies of extractive approach has emerged in the past. All though it's difficult to mention how enough larger interpretative refinement, at sentence or text level add to efficiency. Without natural language processing, the obtained summaries might not be much accurate in terms of semantics. If the input documents cowl multiple topics, it becomes troublesome to come up with a balanced summary. So for more efficient mining of conceptual rule various part of speech other than verb like noun, adjective should also be consider. Semantic analysis can be enhance by including Hypernym and Hyponym. A hybrid Ranking approach can be used to detect more influential sentences with in the document.

REFERENCES

[1] V. M. Navaneethakumar, "Mining Conceptual Rules for Web Document Using Sentence Ranking Conditional Probability", 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME) February 21-22
[2] V.M.Navaneethakumar, Dr.C.Chandrasekar, "A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012
[3] S.Revathi, Dr.T.Nalini, "Clustering and Classification Augmented with Semantic Similarity for Text Mining" , IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013
[4] K.FouziaSulthana, N.Kanya, "Content Extraction with text mining using natural language processing for anatomy based topic summarization" ,International Journal of Modern Engineering Research (IJMER) Vol.3, Issue.1, Jan-Feb. 2013
[5] Claude Pasquier, "Single document keyphrase extraction using sentence clustering and Latent Dirichlet Allocation", International Workshop on Semantic Evaluation, ACL 2010
[6] Amolkumar N. Jadhav1, Ashish Manwatkar2, Deepak Dharrao, "A CONCEPT-BASED MINING MODEL FOR INCREASING TEXT PERFORMANCE" , IJAET/Vol.II/ Issue IV/October-December, 2011
[7] 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
[8] Qiang Zhan, Chunhong Wang, "Hyponymy Extraction Of Domain Ontology Concept Based On Ccrfs And Hierarchy Clustering", International Journal of Web & Semantic Technology (IJWesT) Vol.6, No.3, July
[9] Peter D. Turney, "Coherent Keyphrase Extraction via We Mining", International Joint Conference on Artificial Intelligence (IJCAI-03). August 9-15, 2003.
[10] Bhushan Inje, Ujawla Patil, "Operational Pattern Revealing Technique in Text Mining", 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science
[11] Luepol Pipanmekaporn, Suwatchai Kamolsantiroj, "Mining Relevant Text Features for Retrieving Web Information", 2014 IIAI 3rd International Conference on Advanced Applied Informatics
[12] Ravikarn Punnarut, Gridaphat Sriharee, "A Researcher Expertise Search System using Ontology- Based Data Mining.", Proc. 7th Asia-Pacific Conference on Conceptual Modelling (APCCM 2010).
[13] Xiaoyan Cai, Wenjie Li, You Ouyang, Hong Yan, "Simultaneous Ranking and Clustering of Sentences: A Reinforcement Approach to Multi-Document Summarization", 23rd International Conference on Computational Linguistics Beijing, August 2010
[14] Laxmi Choudhary, Bhawani Shankar Burdak, "Role of Ranking Algorithms for Information Retrieval", nineteenth annual ACM-SIAM symposium on Discrete algorithms; Year: 2008.
[15] Hamid Mousavi, Deirdre Kerr, Markus Iseli, Carlo Zaniolo, "Mining Semantic Structures from Syntactic Structures in Free Text Documents", 2014 IEEE International Conference on Semantic Computing

