Efficient Conceptual Rule Mining on Text Clusters in Web Documents

V.M.Navaneethakumar
Assistant Professor
Department of Computer Applications
K.S.R College of Engineering
Tiruchengode
Tamilnadu
India.

C.Chandrasekar, PhD.
Associate Professor
Department of Computer Science
Periyar University
Salem
Tamilnadu
India.

ABSTRACT

Text mining is a modern and computational approach attempts to determine new, formerly unidentified information by pertaining techniques from normal language processing and data mining. Clustering, one of the conventional data mining techniques is an unsubstantiated learning pattern where clustering techniques attempt to recognize intrinsic groupings of the text documents, so that a set of clusters is formed in which clusters reveal high intra-cluster comparison and low inter-cluster similarity. Most current document clustering methods are based on the Vector Space Model (VSM), which is a widely used data representation for text classification and clustering. Moreover, weighting these features accurately also affects the result of the clustering algorithm substantially. The previous work described the conceptual text clustering to web documents, containing various mark up language formats associated with the documents (term extraction mode). In this work, we are going to present a Conceptual rule mining which is generated for the sentence meaning and related sentences in the document. Weights are appropriated for the sentences having higher contribution to the topic of the document. Conditional probability is evaluated for the sentence weights. Probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed. Experiments are conducted with the web documents extracted from the research repositories to evaluate the efficiency of the proposed efficient conceptual rule mining on text clusters in web documents and compared with an existing Model for Concept Based Clustering and Classification in terms of Topic related rules, Weights of the influential sentence, Topic Sensitivity..

Keywords: Conceptual rule mining, text clustering, conditional probability, probability ratio

1. INTRODUCTION

The modern techniques employed for text mining are concept-based which engage natural language processing in addition to geometric analysis. An association of presented documents and upcoming documents can be completed by the mining functionalities clustering and categorization. It is significant and proficient if the categorization functionality is entrenched into the existing grouping functionality. Approaching to natural language processing, to evade the uncertainty of diverse senses of a distinct word and numerous representation for a distinct sense (depends on the author's vocabulary) NLP can be worned. In the current work a novel synonym based mining model has been projected. It succeeds to all the remuneration of existing concept based mining model. Besides that it savors the spirit of synonym based matching. The current work represents both clustering and categorization at the similar time the work illustrates that the similar parallel measures can be utilized in synonym based approach also.

Text mining specifies to the information drawing out from textual databases or documents. This text mining is different from mining the more types of databases because of its formless form and massive number of proportions. Each word in the document is a dimension. So the primary things for text mining are, giving a construction to the data and dropping the dimensions**. N**ormally in text mining techniques, we calculate the term occurrence of the terms in the document to discover the significance of the term in the document. Alternatively, two terms can contain the similar term occurrence in their documents; so far the meaning abounded by one term is more suitable to the meaning of the sentence than the meaning contributed by the other term.

The Vector Space Model is extensively used document clustering technique and symbolizes data for text categorization and clustering. The terms in the document is symbolized as a feature vector. The terms would be words or phrases. Each feature vector is consigning a term weight supported on the term occurrence of the terms in the documents. Similarity procedures that rely on the feature vector are utilized to discover the correspondence among the documents.

Gradually more, the discovery of a huge number of semantic concepts is being observed as an intermediary step in facilitating semantic video search and reclamation. These semantic concepts wrap a broad assortment of topics that can be approximately classified as objects, sites, proceedings, and specific behaviors and named entities. Researchers have urbanized a huge number of involuntary notion discovery techniques and the most admired approach is to interpret the learning task into numerous binary classification troubles with the subsistence/absence label of every individual concept. Then for each concept, its connected video ideas can be noticed through numerous unimodal classifiers supported audio, visual and text features.

Nevertheless, these binary classification techniques disregard an imperative fact that semantic concepts do not subsist in segregation to each other. They are consistent and associated by their semantic explanations and hence display confident co-occurrence patterns in video collections. For instance, the concept "car" forever co-occurs in a video shot with the notion "road" whereas the concept "office" is not possible to emerge with "road". Such types of concept relationships are not exceptional and it can be probable that mining multi-concept relationship can provide as a practical basis of information to progress the concept detection accuracy.

In this paper, we are going to present a Conceptual rule mining which is generated for the sentence meaning and related sentences in the document. Weights are appropriated for the sentences having higher contribution to the topic of the document. Conditional probability is evaluated for the sentence weights. Probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed.

2. LITERATURE REVIEW

The current techniques that are being utilized for text mining are concept-based which engross ordinary language processing with geometric study. Association of presented documents and imminent documents can be finished by the mining functionalities gathering and categorization. It is imperative and well-organized if the categorization functionality is entrenched into the existing clustering functionality. Imminent to natural language processing, to evade the uncertainty of diverse senses of a distinct word and numerous depiction for a distinct sense (depends on the author's vocabulary) NLP can be worned. In [1], a novel synonym based mining reproduction has been anticipated. It succeeds to all the remuneration of existing concept based mining representation.

Most of the widespread techniques [12] in text mining are supported on the geometric examination of a term, either declaration or phrase or with some connection rule mining method [3]. Statistical study of a term frequency detains the significance of the term contained by a document merely. Nevertheless, two terms can comprise the similar frequency in their documents, but one term supplies more to the denotation of its sentences than the new term. Thus, the fundamental text mining model [2] should point out terms that detain the semantics of text [11].

Text summarization [9] has developed into a significant and appropriate tool for supporting and inferring text information in today's fast-growing in sequence age. It is very tricky for human beings to physically review huge documents of text. Text Summarization [4] is compressing the source text into a shorter version protecting its information contented and in general meaning. It is very complex for human beings to physically review huge documents of text. Text Summarization methods [10] can be divided into extractive and abstractive summarization. An extractive summarization [6] technique comprises of choosing significant sentences, paragraphs etc. from the unique document and concatenating them into shorter structure.

Keyphrases [5] are terms, or set of terms, that detain the key thoughts of a document. They symbolize significant information relating to a document and comprise a substitute, or an accompaniment, to full-text indexing. A relevant keyphrases [7] are also practical to possible readers who can have a rapid summary of the contented of a document and can choose simply which text to interpret. In [8], portray how to excavate association rules in chronological document collections. It explain how to achieve the different steps in the chronological text mining process, counting data cleaning, text alteration, chronological relationship rule pulling out and rule post-processing.

3. PROPOSED CONCEPTUAL RULE MINING ON TEXT CLUSTERS IN WEB DOCUMENTS

The proposed work is efficiently designed for identifying the related concepts in the sentences raised in the web documents based on appropriate weights computed using the Conditional probability. Probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed. The proposed conceptual rule mining on text clusters in web documents is worked under two different phases. The first phase describes the generation of conceptual rule mining for the sentence meaning and related sentences in the document. The weights which have higher contribution to the topic of the document are taken for further process in concept based mining model. The second phase describes the evaluation of conditional probability for sentence weights. Then the Probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed. The architecture diagram of the proposed conceptual rule mining on text clusters in web documents [CRMTC] is shown in fig 3.1.

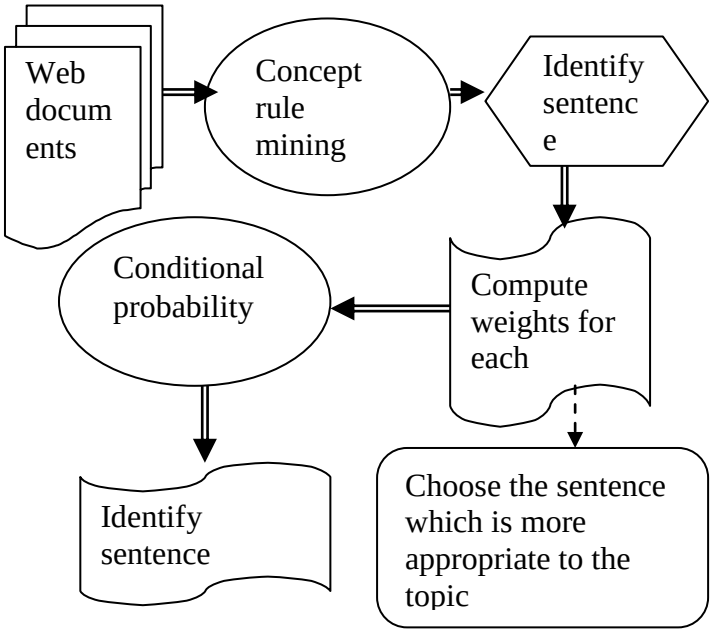

Fig 3.1 Architecture diagram of the proposed CRMTC

From the fig3.1, it is being illustrated that the process of identifying the most appropriate sentence related with the topic given in the web documents. Conceptual rule mining is efficiently created for the sentence meaning and associated sentences in the document. Weights computed for the sentences having higher contribution are taken as the more relevant sentence to the topic of the document. Conditional probability is evaluated for the sentence weights. Probability ratio is recognized for the sentence resemblance from which distinctive sentence meaning causative to the document topic are then listed.

3.1 Generating conceptual rules for web document mining

Conceptual rule mining is generated for identifying the sentence meaning and related sentences in the document. Concepts are a subject of perpetual attention to psychology,

mainly concepts which recognize kinds of things. Such concepts are rational illustrations which facilitate one to differentiate among objects that persuade the concept and those which do not. Roughly all web search engines sustain keyword based exploration, but this style of analysis is powerfully partial in its clarity. The keywords in an analysis do not contain any semantic interconnection except boolean operators. Instances for enviable semantic connections are the subsequent lines: one term is a synonym for a different term or - similar to in the system - one term illustrates a concept or concept land whereas the other one is an occurrence value or a alteration of the first term.

Mining conceptual rules from a web documents can be illustrated as a 3-step process comprising of 1) pre-processing, 2) the actual mining and 3) post-processing. In the pre-processing stage the web documents are renewed from exterior documents into some widespread representation, concepts/words are mined (tokenization), and then different operations might be achieved on the text intending at rising the quality of the results or dropping the running time of the mining process. Then the actual mining is performed, resulting in a number of conceptual rules. The number of rules might be very high, and in the post-processing phase the system attempts to decide which rules are most fascinating, based on some measure.

3.1.1 Pre Processing

In this step the concepts and sentences are mined from the web documents. Depending on the application field, terms that are numeric values may or may not be reserved for mining. This step is a mapping from web documents to a list of list of terms. Web documents can be huge and each comprises a huge number of discrete terms. In order to enhance quality of rules in addition to decrease the computational cost, in common only a separation of the terms in the documents is essentially component of the rule mining process. In the text filtering and refinement step it is dogged which of the terms that shall contribute, and definite transformation may also be achieved so as to enlarge the possibility of constructive rules. This step comprises of a number of processes, each containing as input a list of list of terms and generating a new list of list of terms. The goal of text filtering is to eliminate terms that can be implicit to not throw in to the production of meaningful rules.

3.1.2 Rule mining

In our case, we wish to launch a proposition space of rules, and investigate the behavior of a coherent web documents demanding to study those rules from concept examples. Conceptual rule mining is generated for identifying the sentence meaning and related sentences in the document. Thus the learning problem is to establish P (F|E, λ (E)), where F depicts over rules, E is the set of experiential example concepts (perhaps with repeats) and λ (E) are the observed concepts. This measure may be expressed (through Bayes' formula) as:

$$P(F \mid E,\lambda(E)) \alpha P(F)P(E,\lambda(E) \mid F) \quad \text{… eqn 1}$$

To use this relationship we will require, besides a hypothesis space, the prior possibility, P (F), and a probability function, P (E, λ(E) |F). The hypothesis space of rules is specified by well formed formulae of a concept language, which is précised by a context free grammar over an alphabet of fatal symbols. To obtain a likelihood function, we commence by creating the weak sampling postulation, that the set of experiential examples is self-sufficient of the concept:

$$P(E,\lambda(E) \mid F) = P(\lambda(E) \mid F,E)P(E) \quad \text{…… eqn 2}$$

The term P (E) will terminate when all feature values are experiential for all objects. Next we believe that the concept is true precisely when an object satisfies the hypothesized formula. After identifying the sentence meaning and related sentences in the document, the weights of each concept have to be computed to identify the sentences which are higher contribution to the topic of the document.

To examine every concept at the sentence level, a new concept-based frequency determination, called the conceptual term frequency *ctf* is used [1]. The ctf computations of concept c in sentence s and web document wd are as follows:

At sentence level (conceptual term frequency ctf):

The *ctf* is the number of occurrences of concept c in sentence s. The concept c, which normally emerges in diverse verb case structures of the similar sentence s, has the major role of causative to the meaning of s. In this case, the *ctf* is a limited appraise on the sentence level. A concept c can have several ctf values in diverse sentences in the similar document d. Thus, the ctf value of concept c in document wd is calculated by

$$ctf = \frac{\sum_{n=1}^{sn} ctf}{sn} \quad \text{…….. eqn 3}$$

Where sn is the entire number of sentences contain concept c in document d. By evaluating the common ctf values of c in its wd, measures the significance of concept c to the denotation of sentences in document d. A concept, which has ctf values in most of the sentences in a web document, has a major contribution to the meaning of its sentences that leads to discover the topic of the document. Thus, calculating the average of the ctf values measures the overall importance of each concept to the semantics of a document through the sentences.

At document and corpus level (df,tf):

To examine every concept at the document point, the concept-based term frequency tf, the number of a concept (expression or idiom) c present in the unique document, is designed. The tf is a limited appraise on the document point. To extract concepts that can discriminate between documents, the concept-based document frequency df, the number of documents containing concept c, is calculated. The df is a inclusive appraisal on the corpus point. Based on the frequency at three levels , weightage will be given to each concept. The more significant concept will have more weight. The weights can be calculated as follows.

$$weight_i = (tfweight_i + ctfweight_i) \times \log(\frac{N}{df_i}) \text{ .. eqn 4}$$

$$tfweight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn} (tfij)^2}} \quad \text{……. Eqn 5}$$

$$ctfweight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn} (ctf_{ij})^2}} \quad \text{…… eqn 6}$$

Where $ctfweight_i$ - value identifies the weight of the concept i in document wd at the sentence level

$tfweight_i$ - value identifies the weight of concept i in document wd at the doc. Level

$\log(\frac{N}{df_i})$ - weight of the concept i, when i shows in a small amount of documents cn is the sum of the concepts which has tf value in web document wd.

$tfweight_i + ctfweight_i$ - precise evaluation of the role of each concept to the topics stated in a web document. Through the above process, the sentences which are more relevant are identified related to the topic of the web document.

3.1.3 Rules Post processing

In text mining and temporal text mining in scrupulous, a very huge number of conceptual rules will be established. A very significant and demanding problem is to recognize those that are appealing. Similar to conventional inter-transaction abstract rules, parameters like number of concepts and measures like confidence and support are also significant when generating inter-transaction item sets and choosing final rules. Unlike conventional rule mining where regularly as great as probable item sets are formed, in mining rules in text generally the rules based on comparatively small item sets are sufficient. It should also be declared that as of the number of discrete terms the distinctive minimum support for conceptual rules in text databases can be moderately low.

One meticulous feature of rule mining in text is that regularly a high support means the rule is too evident and thus less appealing. These rules are frequently a result of happening terms and can partially be removed by identifying the suitable stop words. Nevertheless, many will stay, and these can to a definite amount be unconcerned by indicating a maximum support on the rules, i.e., the only resultant rules are those over a definite minimum support and less than a definite maximum support. Another technique that can be used to remove unwanted rules is to specify stop rules, i.e., rules that are widespread and can be removed without human intervention.

3.2. CONDITIONAL PROBABILITY TO IDENTIFY THE SENTENCE SIMILARITY

Assume a web document wd with different set of terms of t1, t2,t3 …, and each terms is presented with prior probability associated with it denoted as P(t). In Conditional probability scenario, the probability of relevance of the sentence to the topic of the web document wd given as,

$$P(t_n \mid q, wd) \approx (P(q \mid wd)) \quad \text{…….. eqn 7}$$

$$P(q \mid wd) = P_{wd}(q)$$

$$= \sum_t P_{wd}(t) I(t,q)$$

$$= \sum_t P(t)(1+\lambda_{wd}) I(t,q) \quad \text{…… eqn 8}$$

Where $P_d(t)$ – Posterior probability of the terms t occurred in web document wd.

λ_d – ratio by which prior probability is altered.

q- query

I (t,q) – 1 if term t presents in query q, 0 otherwise

The value of λ_d is the ratio of sum of probabilities of the terms t not occurred in web document wd and the probabilities of the terms t occurred in web document wd. Conditional probability is evaluated for the sentence weights. Probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed. Through the above equations (eqn 7, 8), the similarity of the sentences are identified and pick the most appropriate relevant sentence related to the topic of the document in a reliable manner. The next section describes the experimental evaluation to estimate the performance of the proposed CRMTC.

4. EXPERIMENTAL EVALAUTION

A widespread of experiments are conducted to check the efficiency of concept matching by evaluating the appropriate weights for the sentences related to the topic in designing a precise fortitude of the comparison among webs documents extracted from the research repositories using the concept rule mining for web document based text clustering. The proposed conceptual rule mining on text clusters in web documents [CRMTC] is efficiently done by concept based mining model. The experimental evaluation tests aimed at comparing the existing efficient concept based mining model for enhancing text clustering with the proposed web document based text clustering using concept based mining model. At first, Conceptual rule mining is generated for the sentence meaning and related sentences in the document. Then, weights are appropriated for the sentences having higher contribution to the topic of the document are analyzed. Conditional probability is evaluated for the sentence weights and probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed. The performance of the proposed conceptual rule mining on text clusters in web documents is measured in terms of

i) Topic related rules

ii) Weights of the influential sentence

iii) Topic Sensitivity

5. RESULTS AND DISCUSSION

Compared to an existing Model for Concept Based Clustering and Classification [MCBC], in this work, we have seen that how related sentences are identified using conceptual rule mining. It described the process of the generating the sentence which is more relevant to the topic and conditional probability is evaluated for the sentence weights. Probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed. The below table and graph described the performance of the proposed conceptual rule mining on text clusters in web documents.

Table 5.1 No. of rules vs. topic generated rules

No. of rules	Generation of Topic related rules	
	Proposed CRMTC	Existing MCBC
2	20	8
4	31	13
6	43	19
8	52	22
10	60	29

The above table (table 5.1) describes the generation of conceptual rules and how efficient the rules are related with the topic in the web document. The results of the proposed conceptual rule mining on text clusters in web documents are compared with an existing Model for Concept Based Clustering and Classification [MCBC].

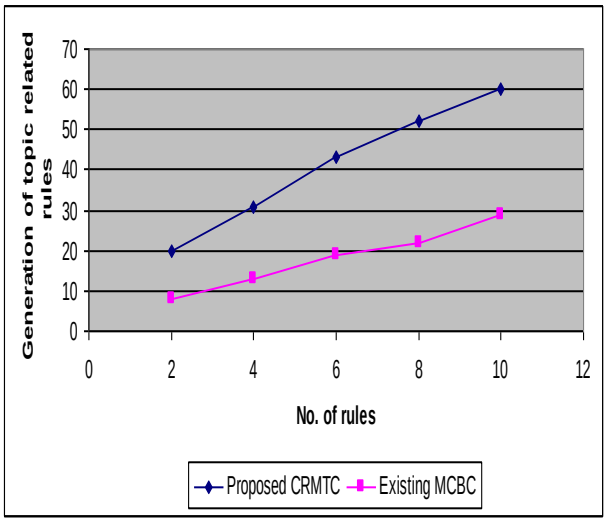

Fig 5.1 No. of rules vs. topic generated rules

Fig 5.1 describes the generation of rules related to the topic in the web document based on the number of rules generated with it. In the proposed CRMTC, the conceptual rule mining is used for generation of rules with respect to concepts in the web document used to identify the sentence meaning and related sentences in the web document. Compared to an existing Model for Concept Based Clustering and Classification which identifies the related sentences and concepts and classified the web documents based on its similarity measures and does not provide an exact result of the relevant sentences for the "web" document, the proposed conceptual rule mining on text clusters in web documents efficiently generate rules based on the concepts / words present in the sentence s for identifying the related sentence to the topic of the web document. The generation of the rules by the proposed CRMTC is more relevant to the topic and the variance is 30-35% high.

Table 5.2 No. of sentences vs. weights of influential sentence

No. of sentences	Weights of influential sentence	
	Proposed CRMTC	Existing SMCBC
10	21	10
20	38	16
30	46	13
40	52	20
50	60	18

The above table (table 5.2) describes the weights of influential sentences related with the topic in the web document. The results of the proposed conceptual rule mining on text clusters in web documents are compared with an existing Model for Concept Based Clustering and Classification [MCBC].

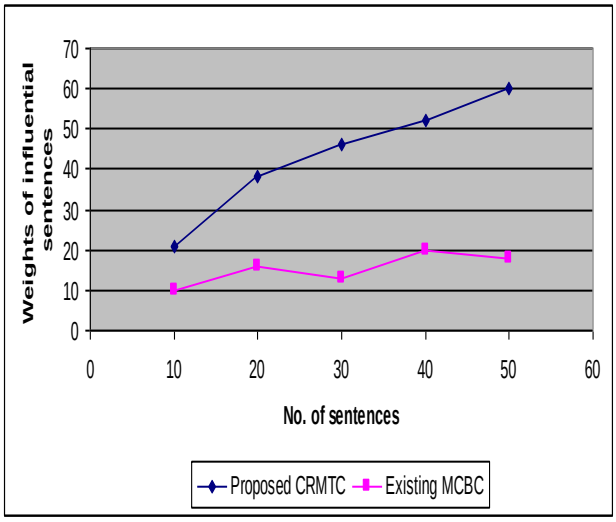

Fig 5.2 No. of sentences vs. weights of influential sentence

Fig 5.2 describes the weights of influential sentences based on the number of sentences present in the web documents and percentage of identifying the more appropriate sentences related with the weights. An existing MCBC identified the significant meaning based on semantic role but not for web documents. The proposed CRMTC used Conceptual rule mining is generated for the sentence meaning and related sentences in the document. Weights are appropriated for the sentences having higher contribution to the topic of the document. Conditional probability has also been evaluated for the sentence weights. Compared to an existing Model for Concept Based Clustering and Classification, the proposed conceptual rule mining on text clusters in web documents provides a good weightage results for the dominant sentences present in the web document which are necessary to identify the relativity of the topics.

Table 5.3 Weights of sentences vs. Topic sensitivity

Weights of sentences	Topic sensitivity	
	Proposed CRMTC	**Existing MCBC**
10	24	10
20	53	30
30	42	26
40	60	19
50	68	34

The above table (table 5.2) describes the sensitivity of the topic in the web document based on weightage of sentences. The results of the proposed conceptual rule mining on text clusters in web documents are compared with an existing Model for Concept Based Clustering and Classification [MCBC].

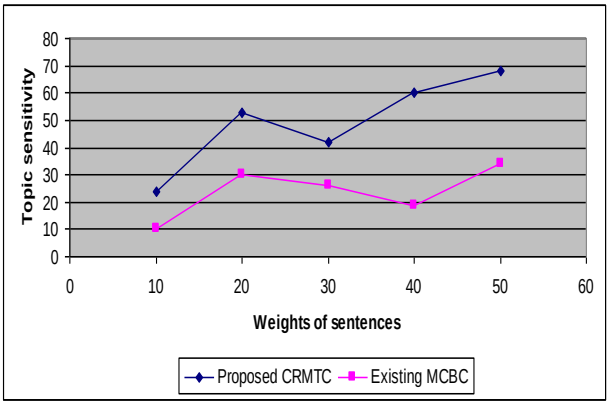

Fig 5.3 Weights of sentences vs. Topic sensitivity

Fig 5.3 describes the understanding of the topic concerning the sentences present in the web document. The proposed CRMTC provides an extension of text clustering related to web documents using conceptual rule mining which done rule pre processing and post processing containing various mark up language formats associated with the web documents (term extraction mode). Based on conceptual rules, the sentence relativity with the topics is identified efficiently and so the sensitivity of the topic is also being high.

Finally, it is being observed that the proposed Conceptual rule mining generated rules for mining the sentence meaning and related sentences in the web document. Weights are appropriated for the sentences, given higher contribution to the topic of the document. Then the Conditional probability is evaluated for the sentence weights. Probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed.

6. CONCLUSION

The proposed work web documents based text clustering using conceptual rule mining model acts as the opening between natural language processes and text mining restraints. A new concept based mining model collected with the appropriate components, is planned to estimate the text clustering eminence. An extension of conceptual text clustering is done to web documents, containing various mark up language formats associated with the documents (term extraction mode). Conceptual rule mining is applied to identify the sentence meaning and related sentences in the document. Weights are appropriated for the sentences having higher contribution to the topic of the document. Conditional probability is evaluated for the sentence weights. Probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed. The concept-based similarity assessment allowed the measuring of the significance of each concept in esteem to the semantics of the sentence, the subject of the document, and the unfairness among documents obtained. The experimental results showed that the proposed conceptual rule mining on text clusters in web documents outperforms well in terms of topic sensitivity, topic related rules contrast to an existing Model for Concept Based Clustering and Classification.

7. REFERENCES

[1] SaiSindhu Bandaru ET. AL., “An Efficient Semantic Model For Concept Based Clustering And Classification”, International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 4 No. 03 March 2012.

[2] Shady Shehata, , Fakhri Karray, 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, October 2010.

[3] Sotiris Kotsiantis, Dimitris Kanellopoulos , “Association Rules Mining: A Recent Overview”, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

[4] Vishal Gupta et. Al., “A Survey of Text Summarization Extractive Techniques”, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010

[5] Claude Pasquier et. Al., “Task 5: Single document keyphrase extraction using sentence clustering and Latent Dirichlet Allocation”, Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 154–157

[6] Hany Mahgoub et. Al., “A Text Mining Technique Using Association Rules Extraction”, International Journal of Information and Mathematical Sciences 4:1 2008

[7] Kjetil Nørv˚ag et. Al., “Semantic-Based Temporal Text-Rule Mining”, Proceeding on 10th International Conference on Computational Linguistics and Intelligent text processing, CICLing ’09, pages 442-455.

[8] K. Nørv˚ag, K.-I. Skogstad, and T. Eriksen. Mining association rules in temporal document collections. In Proceedings of the 16th International Symposium on Methodologies for Intelligent Systems (ISMIS’06), 2006

[9] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[10] Farshad Kyoomarsi, et. Al., “Optimizing Text Summarization Based on Fuzzy Logic”, In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008

[11] Yongzheng, Nur and Evangelos, “Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora”, WIDM’5, 51-57, Bremen Germany,2005

[12] Rene Arnulfo Garcia-Herandez et. Al., “Word Sequence Models for Single Text Summarization”, IEEE,44-48, 2009.