International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

DOCUMENT CLUSTERING AND SUMMARIZATION BASED ON ASSOCIATION RULE MINING FOR DYNAMIC ENVIRONMENT

J.Jayabharathy[1], S. Kanmani[2]

Pondicherry Engineering College, Puducherry - 605014, India e-mail: bharathyhari_raja@yahoo.co.in

Abstract

Document Summarization is a technique, which reduces the size of the documents and gives the outline and crisp information about the given group of documents. This paper introduces a new update summarization algorithm incorporating association rule mining and correlated concept based hierarchical clustering for dynamic environment. In this algorithm, the associated concepts are extracted using Rule mining technique (Generating Association Rules based on Weighting Scheme) and the Correlated concepts (terms and their related terms) are extracted based on concept extraction algorithm. Extracting concepts based on association rule, helps the user to cluster and summarize the similar concept, which in turn improves the quality of the cluster and the created summary. The performance of the hierarchical clustering based update summarization technique is compared with the existing COBWEB (update summarization) algorithm and static summarization algorithms namely; MEAD, CPLN (Centroid, Position, Length and Numerical value) and CPLNVN (Centroid, Position, Length Numerical value and Verb- Noun) considering Precision, Recall and F-measure as performance metrics. Scientific literature and 20 Newsgroups are chosen as the data set for the experiment analysis. The experimental results demonstrate that the proposed algorithm exhibit better performance, compared to the existing algorithms for summarization.

Introduction

Document clustering helps the user to categorize the document corpuses into relevant groups. Automatic grouping of the text documents enables the user to grasp the needed information from the document clusters that are retrieved for the given query and guides the user to track correct and needed information [1]. Also a collection of web documents are retrieved with different degrees of relevancy for the given query. The user has to explore each and every document to locate the desired information; which is both a time consuming and tiring process. To address these issues, researchers have introduced multi-document summarization technique which produces a concise outline of the retrieved documents called summary. The process of summarization has to be focused for static and dynamic environments. Static Summarization is the process of creating summary for a closed set of document collections. Whereas, updation summarization is the process of renewing the created summary whenever the new document is included to the existing cluster/corpus. The main variation is that dynamic summarization considers the documents' temporal relationship. It also analyzes the relationship between the existing information and the emerging information, which is represented in the dynamic version of the content generated as update summary [2]. Furthermore, constructing an adequate model for information that changes dynamically is difficult and also it is not recognized fully.

Most of the document summarization techniques creates summary for a closed set of documents in a batch mode i.e. for static cluster. In general multi-document summarization [3] is of two kinds: extractive summary and abstractive summary. Update summarization is an emerging concept in the Information Retrieval [4] process. The main goal of update summarization is to provide the user with concise and informative summary with dynamic information related to same topic hence saving user time from browsing the web. The summary is created for the clustered documents assuming that the user has prior knowledge about the topic of the summary.

The major challenge in update summarization is to renew the existing summary without compromising the quality of the existing summary. Update summarization system should also monitor the information change periodically over a given time period. Update summarization is relevant for newswire, since a topic in news stories evolve over time and user/reader would only be more interested about new information about that topic [5]. Most of the research works are based on terms and/or synonyms and hypernyms based clustering and summarization. These techniques fail to capture the relevant information from the domain specific document corpuses. Also these techniques create summary only for the static collection of documents.

Analyzing the limitations in the existing methods, the authors have proposed concept based clustering and summarization for static and dynamic environment. Since it is realized for static environment that concept based summarization gives better performance than the term or synonyms and hypernynms, the authors proposed a new summarization algorithm for dynamic environment named Correlated Concept and Association Rule Mining based Update Summarization Technique (CCARMUST). This algorithm applies hierarchical clustering and summarization based on extracting the Correlated concepts (terms and their related terms). Also the associated concepts are mined using GARW mining algorithm for efficient summary generation. The performances of the above proposed algorithms have been compared with COBWEB (update summarization) algorithm and MEAD [24], CPLN [25] and CPLNVN [13] (static summarization) algorithms considering Precision, Recall and F-measure as performance metrics and the results are presented.

The remaining part of this paper is organized as follows. Section 2 reviews related work on static and dynamic document summarization. Section 3, outlines the existing work considered for comparison through experimental analysis. The section 4 presents the detailed description of the new summarization algorithm CCARMUST. In Section 5, the experimental setup and data set descriptions have been discussed, followed by analysis of results. Finally salient conclusions are presented in section 6.

Related Works

J.Jayabharathy et al., [6] proposed a frequent item set based summary generation algorithm for static documents using the sentence features like length, position, centroid noun and also a new feature noun-verb pair. This algorithm is compared with existing MEAD summarization technique using F-measure as evaluation metric. The results show better performance compared to the MEAD technique. Gaurav Aggarwal et al., [7] proposed update summarization algorithm. The process of clustering the existing set of sentences is done using semantic similarity score. To compute the semantic relation the author used WordNet dictionary. Then centroid is calculated for these clusters and an information content score is computed to identify the new and changed sentence in the subsequent set. To update the summary, relevant sentences are chosen by their position in the original document. This technique does not address the word alignment problem and hence leads to inefficient summary.

Gaurav Aggarwal, et al [8] presented a summarization system which cluster sentences together from the old set based on a semantic similarity score. The authors used the centroid of these clusters, along with an information content score, to identify fresh or changed sentences in the subsequent set. These relevant sentences are ordered by their position in the original document and limited to 100 words to generate the update summaries. ShiyanOu, et.al [9] described a concept-based multi-document summarization system by parsing, extracting information and integrating information. The summarization is done in 4 steps: (1) parsing of dissertation abstracts into five standard sections; (2) extracting of research concepts (often operationalized as research variables) and their relationships, the research methods used and the contextual relations from specific sections of the text; (3) integrating similar concepts and relationships between different abstracts; and (4) combining and organizing the different kinds of information using a variable-based framework.

T E Workman and John F Hurdle [10] proposed a Dynamic Summarization of Bibliographic-Based Data. This technique describes the development of a statistically based algorithm known as Combo that automatically summarizes SemRep semantic predications for a topic and a point-of-view in the Semantic MEDLINE model. This model is evaluated against conventional summarization using a previously established reference standard, in the task-based context of secondary genetic database creation. Xuan Li et al. [11] Proposed a new graph-ranking based method, called QCQPSum, i.e. quadratically constrained quadratic programming problem for update summarization. They mainly address the update property as inequality constraints and perform a constrained reinforcement process to determine sentence salient feature. The previous documents act as constraints without directly participating in the reinforcement propagation in current documents

J.Jayabharathy et al., [12] proposed a modified semantic-based model where related terms are extracted as concepts for concept-based document clustering by bisecting k-means algorithm and topic detection method for discovering meaningful labels for the document clusters based on semantic similarity by Testor theory. The proposed method has been compared to the Topic Detection by Clustering Keywords method using F-measure as evaluation metric. J.Jayabharathy et al., [13, 14] proposed a correlation based concept extraction for document clustering and summarization for both static and dynamic environment. The way the documents are represented as traditional vector space model is replaced by the concept vector. Bisecting K-means algorithm is used to cluster the concept vector. The initial summary is created based on our proposed Correlation Based Multi-Document summarization technique. If a new document arrives, the summary has to be updated. This is done by computing the score for each concept based sentence in the documents.

From the critical analysis of published literature, it is inferred that majority of the clustering and summarization techniques are based on term frequencies. Few researchers in the domain of clustering and summarization techniques and annotation tools use synonyms and hypernyms for predicting the concepts. Moreover, the synonyms and Hypernyms are extracted by means of WordNet lexical database [15]. Since scientific literature and many tracks of news documents consist of purely domain-specific technical terms, the performance of synonyms and hypernyms based clustering may not always yield better results. In order to enhance the quality of the cluster for the above mentioned document sets, the present study carries out clustering and summarization of the document using correlated terms and their associated terms. In this regard, a domain- specific dictionary has been developed by the authors to extract the related terms as concepts.

Overview of the Existing Update Summarization Algorithm Considered For Comparative Analysis

The COBWEB algorithm proposed by Ding wang and Tao Li [1] updates the summary as soon as new document enters the document corpuses. This algorithm is purely based on term frequency. The COBWEB algorithm builds a classification tree incrementally by placing the objects into the tree one after another. The COBWEB algorithm traverses the tree from root node to leaf node for inserting an object into the classification tree. The node insertion is done using the following four operations and this operation is selected according to the maximum Category Utility (CU) function value. The heuristic measure called Category Utility (CU) is used as the criterion function to determine the partition in the hierarchy.

(1) Insert: Add the sentences into an existing cluster.
(2) Create: Create a new cluster.
(3) Merge: Join two clusters into a single cluster.
(4) Split: Separate an existing cluster into several clusters.

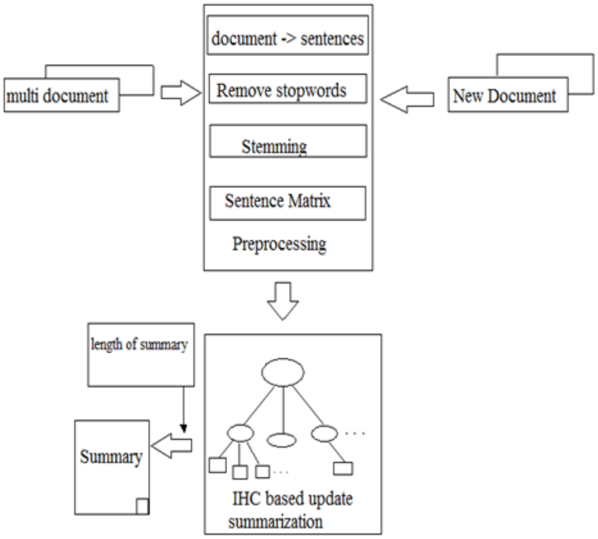

Figure 1. Existing COBWEB System Architecture

Figure 1 demonstrates the framework of the existing COBWEB approach. During the first phase of preprocessing, the documents are tokenized and then the stop words are identified and removed. Typically; the stemming process is performed to transform the words into their rootform. Later, the sentence matrix is constructed for the document(s). During the second phase, the sentences are constructed as tree using COBWEB algorithm. The system generates a sentence hierarchical tree to demonstrate the complete structure of the documents. Segmenting the hierarchy tree at one layer, the COBWEB algorithm creates the summary with the corresponding length at that level onwards.

A. Demerits in COBWEB Algorithm

The quality of clusters and the summary are degraded, as the existing algorithm considers term alone for clustering as well for summarization. This method is more suitable for documents that are related to technical and scientific topics. The process of segmenting the tree at any level for summary generation would not give complete summary of the collected documents and may also lead to redundant information.

Proposed Work

To address the issues in COBWEB algorithm, this paper presents a hierarchical clustering (which is appropriate for clustering the documents incrementally) and topic oriented

ISSN:2319-7900

summarization based concept extraction (which improves the concurrency of the summary). These concepts are extracted using the GARW mining algorithm and correlated concept extraction [12] algorithm for efficient summary generation in dynamic environment. Extraction of the Correlated concepts (terms and related terms) and their associated concepts help to conceive technically related and important information from the domain specific document sources. The proposed topic oriented summarization creates summary according to the given topic, thus the sentences in the summary are more relevant to each other which improves the coherence of the generated summary.

Figure 2 demonstrates the system architecture of the proposed approach.

The overall process has been divided into four major modules.

i) Pre-processing phase: In this phase the document sentences are decomposed using Standford Tokenizer [http:// nlp. stanford. edu/ software/ tokenizer], followed by stop word removal and stemming [16]. Finally, top terms are identified and extracted.

ii) Association Rule Mining (ARM) and Concept Extraction phase: Using the GARW [19] algorithm on the top terms, the rules and the associated terms are extracted. The correlated concepts (terms and their related term) extraction algorithms discussed in [12] is used to identify the concepts. These extracted concepts are referred as Correlated Concept Vector (crtv). The sentences with these concepts are identified and extracted from the document(s) and is given for the clustering phase for tree construction.

iii) Clustering phase: Concepts are clustered using COBWEB based hierarchical clustering. The main reason for using incremental hierarchical clustering based summarization method is to efficiently create updates to the summaries when the document(s) enters the corpuses dynamically. The hierarchy and the selected concepts clearly state the structure of the concepts in the document(s). Wherein, the existing COBWEB algorithm discussed in [17,18] use term/sentence(s) for hierarchical tree construction instead of concepts representation. This in turn improves the efficiency of the algorithm.

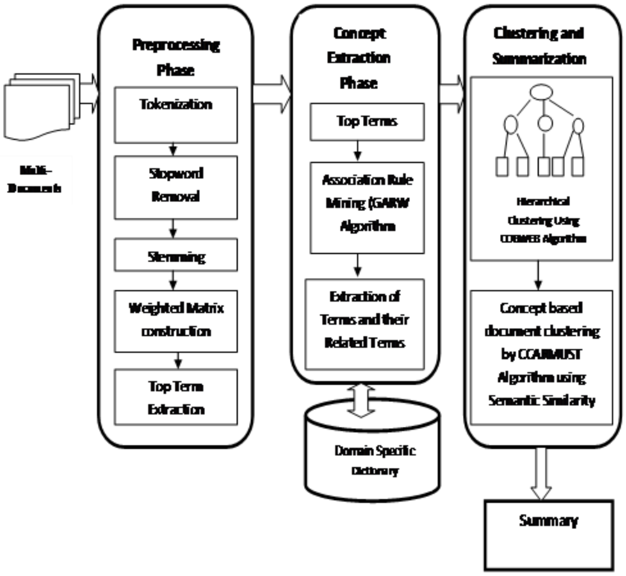
Multi-Documents
Preprocessing Phase
Tokenization
Stopword Removal
Stemming
Weighted Matrix construction
Top Term Extraction
Concept Extraction Phase
Top Terms
Association Rule Mining (GARW Algorithm
Extraction of Terms and their Related Terms
Domain Specific Dictionary
Clustering and Summarization
Hierarchical Clustering Using COBWEB Algorithm
Concept based document clustering by CCARMUST Algorithm using Semantic Similarity
Summary

Figure 2. Proposed CCARMUST System Architecture

iv) Update Summarization phase: Each node in the tree represents the most prominent concepts of the document(s). The most representative concept sentences are selected to form the summary. The users can cut the hierarchy tree at any level to obtain a summary from that height. For example, a user can determine the cutting level based on the length requirement of the summary [1].

A. Why Correlated Terms?

There are many existing clustering algorithms that take synonyms and hypernyms for vector representation. In this study, the authors have considered crtv as concepts for clustering to improve the efficiency of clustering the documents both statically and dynamically. The idea of considering terms and related terms as concepts based on semantic similarity has been carried out for extracting topic from the clustered documents [14]. The proposed technique CCARMUST takes this idea of considering crtv as concepts for clustering and summarization. Considering terms or synonyms and hypernyms for information extraction results to the following issues:

Case 1: Words have multiple meanings, hence diversifies the information extraction.

E.g. Bat : represents the cricket bat or a kind of a bird.

Whereas using correlated concept extraction algorithm the term 'bat' is related to the domian what it refers to.

Case 2: Considering terms or synonyms of the terms limits the search space of the domain.

E.g. wireless: first sense medium of communication.

Mutiple terms equivalent to the term "wireless" is extracted as correlated concepts instead of restricting to one term.

Example, synonyms of the term "wireless" is extracted from WordNet as: "first sense medium of communication", whereas, taking related terms like "wireless", "communication", "protocol" "mobile communication" etc. will be extracted as concepts, which gives better accuracy and improves the efficiency of information extraction. For example, sports article contains terms like: a ball, bat, wicket, run, batsman, over etc. Taking synonyms/hypernyms as concept, will not give better performance since the meaning of these terms are not literally same. If we consider the technically related terms i.e. crtv, all the above mentioned terms will be grouped together as a single concept which refers sports related to the concept – cricket. Similarly the synonym for the term "farmer" from WordNet is extracted as: "a person Title who operates a farm". But, by using the proposed model the concept will be extracted as "farmer", "crops", "fertilizer", "land" and "farm". Clustering the document using this extraction procedure would improve the performance of the resulting cluster, than that of the cluster generated by existing works.

B. The need for Association Rule Mining (ARM)

ARM paves a way for finding information from a collection of indexed documents by automatically extracting association rules from them. Given a set of keywords A= {R1, R2... Rn} and a collection of indexed documents D = {do1, do2... dom}, where each document doi is a set of keywords such that doi⊆A. Let Ri be a set of keywords. A document doi is said to contain Ri if and only if Ri⊆doi . An association rule is an implication of the form Ri⇒Rj whereRi⊂A, Rj⊂A and Ri∩Rj =Ø. There are two important basic measures for association rules, support(referred as s) and confidence(referred as c). The rule Ri⇒Rj has support s in the collection of documents D if s% of documents in D contains Ri∪Rj. The support is calculated by the following formula:

$$Support\left(R_i, R_j\right) = \frac{Support\ Count\ of(R_i, R_j)}{Total\ no.of\ documents\ (D)} \quad (1)$$

The rule Ri⇒Rj holds in the compilation of documents D with confidence c if among those documents that have Ri, c% of them contain Rj also. The confidence is intended by the following formula:

$$Confidence\left(R_i/R_j\right) = \frac{Support\ (R_i, R_j)}{Support(R_i)} \quad (2)$$

From the example, for the term computer networks, (Intranet, Ethernet, LAN, WAN, topology, protocol, Architecture, OSI layers...) are extracted as associated terms. The concept extraction algorithm predicts Network Security, Qos, etc. as their related terms. Extracted terms, related terms and terms inferred through association rule are denoted as correlated concepts term vector (crtv). Clustering based on above mentioned concepts helps to improve the quality of the cluster and the summary in turn.
{Computer networks} => Intranet, Ethernet, LAN, WAN, cluster topology, protocol, Architecture, OSI layers (Association Rule) + QoS, Security (Related terms).

The Steps in GARW algorithm is as follows:

1. Let N denote the number of top keywords that suit the threshold weight value.
2. First, min support is applied to discover all frequent item sets using the top keyword.
3. Find all keywords that satisfy the threshold minimum support. These keywords are called large frequent 1-keyword set 1 *L*.
4. In *k* ≥ 2, the candidate keywords *Ck* of size *k* are generated from large frequent (*k*-1)-keyword sets, *Lk*−1 that is generated in the last step.
5. Scan the index file, and compute the frequency of candidate Keyword sets *Ck* that generated in step 4.
6. Compare the frequencies of candidate keyword sets with minimum support.
7. Large frequent *k*-keyword sets *k L,* which satisfies the minimum support, is found from step 6.
8. For each frequent keyword set, find all the association rules that satisfy the threshold minimum confidence.

C. Hierarchical Clustering Algorithms

This work follows the Katz's distribution based COBWEB algorithm [20] to create hierarchical tree incrementally.

Representative concept Selection for Each Node of the Hierarchy

The sentences from each document, which matches with the concepts *crtv* are identified and extracted, each concept sentence forms a leaf node. If a new concept enters the tree, concept hierarchy is updated by the four operations and also, the representative concepts for the modified nodes are dynamically updated in the following way.

Case 1: For inserting a concept into cluster k, then recalculate the delegate concept R_c of cluster k using

$$Rc = ArgmaxC_i((1-\alpha)SSim\,(query, C_i) + {}^{\alpha}/_{K}\,\Sigma SSim(C_i, C_j)) \quad (3)$$

I≠j Where K is the number of concepts in the cluster and SSims() is the Semantic similarity function between pair of concept.

Case 2: For creating a new cluster k, the newly in coming Concept K_{new} represents the new cluster, i.e., A_{rc}=K_{new}.

Case 3: For merging two clusters cluster A and cluster B into a new cluster C, the concept obtaining the higher similarity with the query is selected as the representative concept at the new merged node.

$$Rc = Argmax\,R_a,R_c\,((SSim(query,R_a),SSim(query,R_c)) \quad (4)$$

Case 4: For dividing the cluster A into a set of clusters as {cluster1, cluster2... cluster n}, take away node ‘a’ and alternate it using the roots of its sub tree. The matching delegate concepts are the delegate Concept for the unique sub tree roots {R1, R2... Rn}.

This algorithm uses semantic similarity for identifying the similarity between two concept sentences. The semantic similarities of the two sentences is computed based on overall score computation discussed in [21], wherein the existing algorithm use cosine similarity measure to compute the similarity between the sentences. When all the documents/concept is represented as a hierarchy tree, this tree clearly shows the structure of the texts. The summarization process is discussed in the following section.

D. Summarization

In the proposed work the summarization modules considers the clustered documents as the input and generates accurate information as summary. In the existing update summarization, user can cut the hierarchy tree at their desired layer to get a summary from that height. For example, a user can determine the cutting level based on the length requirement of the summary discussed in [1]. In this paper the authors proposed a new method of summarization which is based on given topic. The procedure involved in summarization is:

i) Get the topic and desired percentage (length of the summary) from the user for the summary creation.

ii)Generate the concepts for the given topic, named as tcv = (tc1, tc2 ….tcn) based on association rule mining and correlated concept extraction algorithm.

iii) For each tci where i = 1 to n.

iv) Identify the root node for which the topic tci matches, for every leaf node of the sub tree, compute semantic similarity between the topic tci and the leaf node (concept sentence). The node which has maximum similarity match with the given tci is extracted and added to the summary.

v) Extract all the sentences which match with tci. Let S1, S2 …..Sm be the set of sentences extracted.

vi) Compute redundancy penalty (discussed in next section) between the sentences in the summary and the new sentence Si.

vii) Include the redundancy penalty along with the sentence Si.

viii) The sentences with less penalty score are included to the summary.

Concept Based Redundancy Elimination

Captions In extraction based summary, including top ranked sentences from all the documents would lead to redundancy which in turn leads to inefficient summary. Hence when taking a top ranked sentence to be included in the summary, it is first verified whether the same sentence is already present in the summary or not. If the sentence already exists, then the sentence gets redundancy penalty and is discarded from adding it to the summary. If not, the sentence is considered for summary creation. The elimination of redundant sentences using correlated concepts gives efficient results. The reason for this efficiency is the words in the sentences are synonymously different but computing concept wise commonality will lead to redundant sentences, hence achieves better results. The algorithm for Concept based Redundancy Elimination Technique (CBRET [14] is given below:

```
Procedure CBRET(S_i, S_j,TV)
begin
  Input:
  •  Let $S_i$ and $S_j$ be the sentences for which redundancy is to be computed
  •  Assume $S_i$ comprises of n terms and $S_j$ consists of m terms excluding the stop words
  •  Let Tv be the threshold value assigned as 0.5

  for l = 1 to n.
  $C_i$ = Extract concepts for the term $t_l$ in $S_i$ based on the algorithm discussed in section 3.1
```

```
for k = 1 to m
        $C_j$ = Extract concepts for the term $t_k$ in $S_j$ based on the algorithm discussed in section 4.1
// Let $C_i$ and $C_j$ are the list of extracted concepts of $S_i$ and $S_j$.
// Compute the redundancy penalty. The formula for calculating redundancy penalty is:

    $R_s = \frac{2*(No.of\ overlapping\ concepts)}{(No.of\ concepts\ in\ C_i + No.of\ concepts\ in\ C_j)}$        (5)

If ($R_s$= 1) then it is assumed as identical sentences
        If (Rs = 0) there are no common words (concepts) in the sentences $S_i$ and $S_j$

If ($R_s$>TV)

begin

    • Then the sentence gets the redundancy penalty.

    • Subtract the redundancy penalty score $R_s$ from the original sentence score computed based on sentence features.
end for.
end CBRET.
```

Concept Based Sentence Ordering Technique

Sentence reordering is an important concept in summarization because, taking sentences from many documents and including in the summary as such will not give proper meaning and reduce the flow of readability. The proposed sentence ordering algorithm is also based on concepts extraction. The similarity between the sentences calculated during redundancy elimination step is used for sentence ordering. If the similarity between the concepts of the two sentences is high but less than 0.5 will be placed as subsequent sentences. (If similarity is more than 0.5, the sentence will get redundancy penalty and gets eliminated).

Experimental Results

The dataset used for our experimental setup contains 500 abstract articles collected from the science direct digital library. These articles are classified according to the science direct classification systems into 5 major categories: computer networks, high speed networks, quantum mechanics, essential physics, and mobile communication network. For newsgroup: 2G spectrum, flight crash, politics, terrorism articles are gathered from 20 newsgroups.

A. Performance Metrics

i) F-measure [24, 25]: F-measure combines the precision and recall from information retrieval. Each cluster is treated as if it were the result of a query and each class as if it were the desired set of documents for a query. The recall and precision of that cluster for each given class are calculated.

The precision and Recall is measured using the following parameters.

- **Correct**-the number of sentences extracted by the System as well as by the human
- **Wrong**-the number of sentences extracted by the system but not by the human
- **Missed**-the number of sentences extracted by the human but not by the system

Precision and Recall is computed as:

$$\text{Precision} = \frac{\text{Correct}}{\text{Correct+Wrong}} \quad (6)$$

$$\text{Recall} = \frac{\text{Correct}}{\text{Correct + Missed}} \quad (7)$$

F-Measure is computed by using the formula given below

$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision + Recall}} \quad (8)$$

- Precision reflects correctness of number of systems extracted sentences.
- Recall reflects number of missed sentences by the system.
- F-Measure ranges from 0 to 1.

B. Implementation Procedure

Initially, text documents which have been collected from various sources were accumulated in a database. Then, preprocessing was carried out by considering the various stages like: tagging by means of Stanford POS tagger tool, stop word removal and stemming using Porter Stemmer algorithm and morphological capabilities of WordNet. The above preprocessing is common for both existing and proposed algorithms considered in this study. In the existing

ISSN:2319-7900

work, documents are represented as sentence matrix and it is clustered using COBWEB algorithm. For implementing the proposed algorithm as discussed in section 4 is applied, along with dataset chosen for the study were used. On applying GARW and concept extraction algorithm, concept vector crtv is formed; this sentence vector is used for constructing the hierarchical tree. The COBWEB, MEAD, CPLN and CPLNVN algorithms as originally proposed by the various authors were implemented in experimental setup. The proposed CCARMUST algorithm is implemented as discussed in section 4 and the results are analyzed and it is discussed in the following section.

C. Results and Comparative Analysis

The performance analysis of the existing COBWEB based update summarization and the proposed CCARMUST algorithm are categorized into two classes:

i) Based on newsgroup dataset.
ii) Based on Scientific Literature

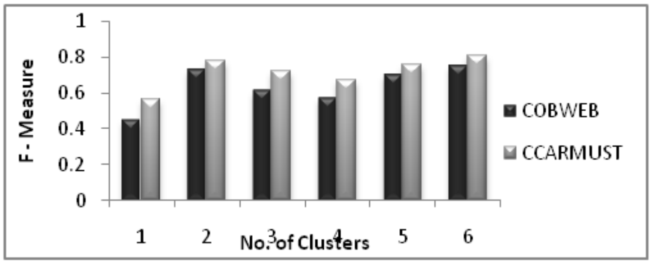

Figure 3. Comparison of document update summary for newsgroup using F-measure

Figure 3 and 4 shows the performance of F-Measure for news group and scientific dataset respectively by considering update summarization using COBWEB algorithm and the proposed Correlated Concept and Association rule mining based Update Summarization Algorithm. The overall F-measure quality of proposed technique has an improvement of +6.34% to +11.21% and +3.32% to +5.5% against COBWEB algorithm for Scientific and Newsgroup dataset respectively.

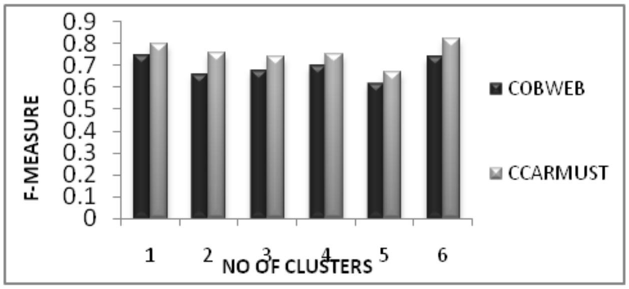

Figure 4. Comparison of document update summary for scientific data set using F-measure

To evaluate the quality of the updated summary, experiments were conducted by implementing the existing term based static summarization techniques like MEAD[24], CPLN [25] and CPLNVN [13] for the scientific and newsgroup data set, considering the same set of document clusters which includes the newly arrived documents also. The documents are processed and clustered using Bi-secting K-means algorithm and summary is created based on MEAD, CPLN and CPLNVN techniques. The created summary are evaluated and analyzed against COBWEB and CCARMUST algorithms and results are presented in figure 5 and 6.

A comparative analysis of the proposed and existing COBWEB algorithm has been done. The performance of the proposed CCARMUST algorithm gives better results compared to the existing COBWEB algorithm. This is because the data set chosen for these experiments are domain specific documents which consists of more scientific and technical terms compared to English literary terms. Also, we have made a performance evaluation with respect to each data set. From Figure 5 and 6, it is clear that the summary based on correlated terms with redundancy elimination and concept based sentence ordering gives better performance than the existing COBWEB algorithm for scientific documents.

ii)Cluster Coherence Comparison: To justify the quality of the clusters formulated on applying the existing and the proposed clustering algorithm, the authors analyzed the cluster coherence using Purity metric because the quality of the summary hardly depends on the coherence of the cluster. The authors also made an attempt in the proposed algorithm by applying correlated concept extraction algorithm first and then GARW algorithm for mining the concepts. This trail doesn't yield better results because, concept algorithm extracts the maximum number of concepts from the document collection and thus applying GARW in the later stage becomes ineffective.

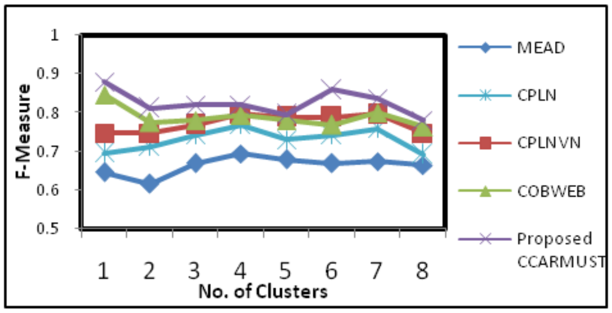

Figure 5. Performance comparison of static and update summarization algorithms for Scientific Literature Dataset

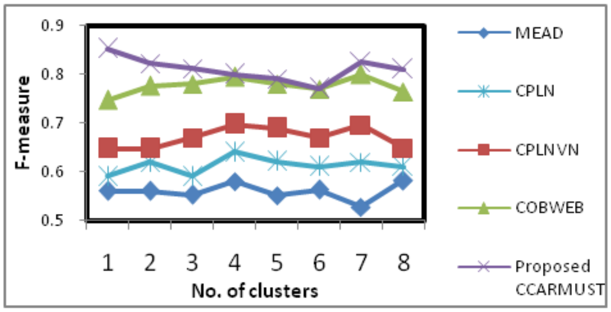

Figure 6. Performance comparison of static and update summarization algorithms for Newsgroup Dataset

iii) Purity [22]: The purity measure evaluates the coherence of a cluster, that is, the degree to which a cluster contains documents from a single class. Given a particular cluster Ci of size ni the purity of Ci is formally defined as

$$P(C_i) = \frac{1}{n} \max(n_i^h) \tag{9}$$

Where max (nih) is the number of documents that are from the dominant class in cluster Ci and nih represents the number of documents from cluster Ci assigned to class h. The overall purity of a clustering solution is:

$$Purity(S) = \frac{1}{n} \sum_{i-1}^{n} \max(n_i^h) \tag{10}$$

Figure 7 and 8 shows the results analysis of Purity metric for scientific documents and newsgroup respectively by considering hierarchical clustering using COBWEB algorithm and the proposed Correlated Concept and Association rule mining based Update summarization. From figure 6 and 7 it clearly states that the clusters formed through extracted concepts shows better results than that of term based COBWEB.

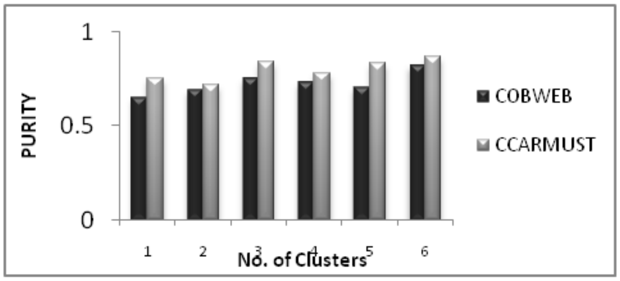

Figure 7. Comparison of cluster coherence for Newsgroup dataset

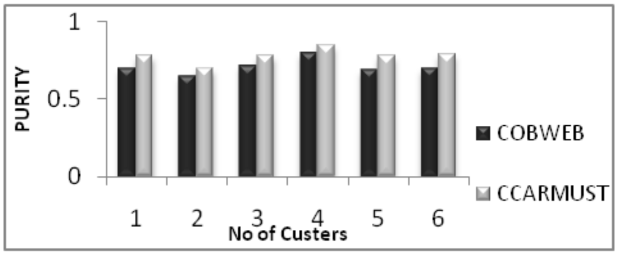

Figure 8. Comparison of cluster coherence for scientific dataset

Conclusion

Dynamic collection of information has become essential in today's life to keep the information updated. Hence, there arises the need for updating the existing summary frequently. Our proposed work, updates the summary efficiently based on topic concepts. The proposed technique also concentrates on redundancy elimination and sentence ordering in the summary which leads to an effective readable summary. The results prove that our proposed technique outperforms the existing COBWEB algorithm. To judge the quality of the created summary, comparative analysis between the static and update summarization algorithms have also been attempted and the results shows better performance as the proposed algorithm has better coherence as we adopt concept based clustering, summarization, redundancy elimination and sentence reordering.

Future Enhancements: In the future, the system could be enhanced to identify topic from the summary. Also this could be extended to other domains. This work mainly concentrates on inclusion of new documents to the cluster and update summary, deletion of documents and renewal of summary according to deletions could also be addressed. In future, instead of using GARW algorithm other equivalent algorithms can be experimented on fuzzy association rule mining for more accurate results and the system could be enhanced to summarize not only text documents but also other type of documents like PDF etc.

References

[1] A Dingding Wang, Tao Li. "Document Updates Summarization Using Incremental Hierarchical Clustering", In the Proceedings of CIKM '10 Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ISBN: 978-1-4503-0099-5, pp 279-288, Oct 2010.

[2] Mei-Ling LIU, De-QuanZheng, Tie-Jun Zhao Hong-E RenAndyang YU , 'Dynamic Multi-Document Summarization Research based on Matrix Subspace

Analysis Model', Journal of Information and Computing Science, Published by World Academic Press, Vol. 6, No. 3, pp.227-233, 2011.

[3] Jade Goldstein, Vibhu Mittal, Mark Kantrowitz and Jaime Carbonell , 'Multi-Document Summarization by Sentence Extraction, Automatic Summarization', In Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, Seattle, Washington, April 30, pp 40-48, 2000.

[4] Singhal and Amit,"Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): pp35–43, August 2001.

[5] Jin Zhang, Xueqi Cheng, HongboXu, XiaoleiWang, YilingZeng, 'Summarizing Dynamic Information with Signature Terms Based Content Filtering', ICT-CAS's: TAC Document Understanding Conference, National Institute of Standard and Technology, http://duc.nist.gov/, November 17-19, . Maryland, USA, 2008.

[6] J.Jayabharathy, S.kanmani and T. Buvana "Multi-document Summarization based on Sentence Features and FrequentItemsets" In the Proceedings of the Second International Conference on Computer Science, Engineering and Applications (ICCSEA 2012), Volume 1, pp 657-671, 2012.

[7] Gaurav Aggarwal, RoshanSumbaly, Shakti Sinha, "Update Summarization", Communications of the ACM, Vol. 48, No. 10, April 2009.

[8] Gaurav Aggarwal, "Semantic Similarity", Communications of the ACM, Vol. 60, No. 22, April 2008.

[9] Shiyan Ou, Christopher S.G. Khoo and Dion H. Goh, "Design and development of a concept-based multi-document summarization system for research abstracts", Journal of Information Science OnlineFirst, published on December 3, 2007.

[10] T E Workman and John F Hurdle Dynamic summarization of bibliographic-based data. Available at: http://www.biomedcentral.com/1472-6947/11/6.

[11] Xuan Li, Liang Du and Yi-Dong Shen ,'Update Summarization via Graph-Based Sentence Ranking', IEEE Transactions On Knowledge And Data Engineering", ,Vol PP, issue 99, pp 1-14.2012

[12] J. Jayabharathy, S. Kanmani and A. AyeshaaParveen, "Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature", 2nd International Conference on Data Storage and Data Engineering, DSDE2011, 13th-15th May 2011,China.

[13] Jayabharathy, S.Kanmani, N.Sivaranjani, "Correlation Based Multi-Document Summarization for Scientific Articles and News Group", International Conference on Advances in Computing, Communications and Informatics – ICCACI'12 In Proceedings in ACM, 3- 5 August, Chennai, India, pp 1093-1099,2012.

[14] Jayabharathy, J. and Kanmani,S. Multi-Document Update Summarization Using Co-related Terms for Scientific Articles and News Group, , Int. J. Computational Science and Engineering, Inderscience (Accepted for Publication, Article In Press),2013

[15] G. A. Miller .WordNet: A Lexical Database for English, Communication. ACM, 38(11) pp 39-41,1995.

[16] WB Frakes, CJ. Fox,.Strength and Similarity of Affix Removal Stemming Algorithms. ACMSIGIR Forum, pp 26-30, 2003

[17] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, pp 139–172, 1987.

[18] J. H. Gennari, P. Langley, and D. Fisher. Models of incremental concept formation. Journal of Artificial Intelligence, pp 11–61, 1989.

[19] Hany Mahgoub"A Text mining technique using Association Rules Extraction" In the International Journal of Information and Mathematical Sciences, vol (4), August 2008.

[20] N. Sahoo, J. Callan, R. Krishnan, G. Duncan, and R. Padman. Incremental hierarchical clustering of text documents. In Proceedings of CIKM 2006

[21] Walaa K. Gad, Mohamed S. Kamel,. Incremental Clustering Algorithm Based on Phrase- Semantic Similarity Histogram. Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, 11(14), pp 2088-2093, 2010.

[22] Anna Huang, "Similarity Measures for Text Document Clustering", New Zealand Computer Science Research Student Conference (NZCSRSC), (14-18 April 2008), pp. 49-56, April 2008.

[23] http://en.wikipedia.org/wiki/Precision_and_recall# Precision.

[24] Dragomir Radev, Jahna Otterbacher, Adam Winkel, and Sasha Blair-Goldensohn, NewsInEssence: Summarizing Online News Topics, Communications of the ACM, Vol. 48, No. 10, October, pp 95 – 98, 2005

[25] Kogilavani, A. and P. Balasubramanie , 'Sentence Annotation based Enhanced Semantic Summary Generation from Multiple Documents', American Journal of Applied Sciences Vol 9 Issue 7: pp 1063-1070,2012.