Journal of Theoretical and Applied Information Technology
20th August 2015. Vol.78. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RANK BASED DOCUMENT CLUSTERING AND SUMMARIZATION APPROACH IN THE DISTRIBUTED P2P NETWORK

[1] A.SRINIVASA RAO, [2] Dr.Ch.DIVAKAR, [3] Dr.A.GOVARDHAN

[1] Research Scholar,Dept of CSE JNTUK Kakinada, A.P ,India.
[2] Principal, Pydah college of engineering and technology, Visakhapatnam,AP,India.
3 Professor and Director, School of Information Technology, JNTUH,Hyderabad,Telangana state.

ABSTRACT

Quick and high quality document clustering techniques play a vital role in text mining applications by grouping large text documents into meaningful clusters and enhancing the clustering accuracy using dimensionality reduction or query expansion. Detecting meaningful clusters and summaries in Distributed p2p network applies single document summarization techniques and peer relationships for detecting meaningful clusters and summaries. Traditional cluster based summarization methods usually suffer with the computation speed, compression, peer selection and sentence clustering in order to generate high quality summaries. Traditional document clustering and summarization methods assume node adjacency and neighborhood information to build clusters and summaries. Since the multilevel overlay p2p networks have suffered with node adjacency and duplicate information, it was difficult to generate optimal clusters and summaries within the peers. Proposed approach provides better solution to generate optimal document clustering using probabilistic k- representative clustering algorithm and forms efficient summaries using phrase rank based summarization. Experimental results give better performance in terms of execution time, entropy and cluster quality are concerned.

Keywords: *Document clustering, Summarization, Phrase ranking,P2P networks.*

1. INTRODUCTION

Document classification and clustering have been assessed as a document retrieval and visualization approach [3]. Document summarization and clustering techniques attempt to cluster documents simultaneously based upon their common characteristics; the documents that are significant to a certain concept will optimistically assign in a unique cluster [4]. For new guests or users, an efficient categorical or nominal index of the whole forum as well as an online searching service will assist them to look for their user specific cluster sentences. For the skilled members in p2p networks, an automatic peer document clustering will assist them to assess and determine high quality documents more efficiently and easier.The goal and purpose of the work is to enhance a document summarizing using clustering methods to group the web documents in an online distributed networks. Such a summarization result has definitely boosted up the knowledge relationship in the distributed p2p network. Any clustering procedure depends on 4 principles:

(1) Data illustration model,
(2) Document similarity measure,
(3) Cluster development system, and
(4) Clustering method that develops the clusters/summaries using the boosting model and the phrase similarity.

The majority of the document clustering techniques that are in use today are centered on the feature vector spaces[1, 2], which are broadly used to train document model for text clustering and classification. Each featured vector space specifies documents as a characteristic vector of the terms that occur in all the document collection set. Each document feature vector includes word/ phrase frequencies of the words appearing in that sentence or document.Document Similarity between sentences/documents are examined using one of document similarity measures that are based on such a feature vector or word frequencies. For instance, Jaccard measure and the cosine measure . Clustering techniques based on this vector spaces make use of single word i.e one gram interpretation only, they do not make use of any word neighborhood or phrase based clustering.

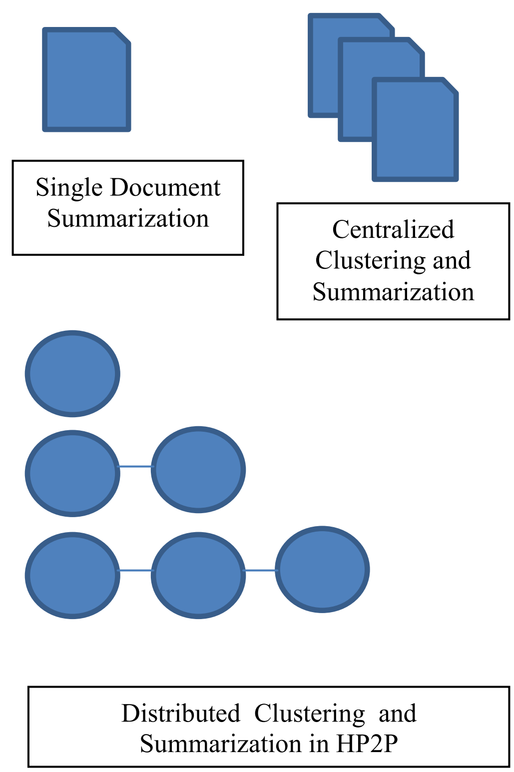
Journal of Theoretical and Applied Information Technology
20th August 2015. Vol.78. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Document summary presents the user with topic of the original documents is often called a document summary. A summary that can be read in place of the document is called an informative summary. An informative summary will include facts that are indicated in the input document or set of sentences, while an indicative summary may provide capabilities such as writing style, length etc.
However, clustering can be executed to many kinds of data, the focus of this literature is on clustering p2p documents [4] which is a sub-field of text-mining. Document clustering deals with the unsupervised learning of a document selection in valuable groups based on their textual content, typically for the determination of topic identification; i.e. documents in one cluster belongs to a particular topic, while different clusters symbolize different topics or sentences. Unlike document classification – which is a supervised method that expects prior information about document groups to train a classification, document segmentation is an unsupervised learning technique that does not depend upon prior identification of topics.

The main usage, we take into consideration here is the guidance of a p2p based virtual node extended supercomputer for equivalent calculations in the parallel synchronous process [8] . The bulk synchronous parallel process provides the outline view of the practical arrangement and the connection capabilities of the network hardware (e. g., a cluster of workstations,a parallel computer or a set of nodes connected by the wireless network). A bulk synchronous parallel system comprises of a set of bulk procedures and a series of super-steps with bounded time periods of a level synchronization. Inside a super-step each procedure operates internal calculations and transmits information to other external procedures; soon after it signifies, by calling the sync procedure, that it is set for the block synchronization.
A distributed type of this method is also used in the hierarchical distributed clustering method to generate summaries for the consistently distributed clusters [5]. Figure 1, demonstrates the different levels at which summarization of document clusters can be performed.

- Key-phrase extraction can be carried out to a single document for tagging the document/sentence; this is generally used in metadata (e.g. Description, title, keywords) that can be correlated with the document.
- A consolidated document based clustering can be labeled and summarized using key feature phrase extraction.
- Extended document clusters or classifications in a flat p2p network can be optimally summarized.



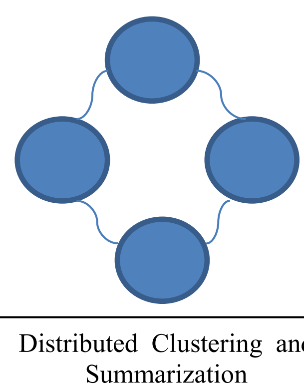


Figure 1. Single Document Summarization Process

Document Cluster summaries can be interchanged between peers to assist cooperative clustering.
- Distributed clustering in a hierarchical network can be summarized level by level using specified threshold. Cluster phrase summaries can be

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

operated at different heights of the hierarchy, thus featuring different types of summaries ranging from general to broad level.

2. RELATED WORK

Various documents clustering techniques have been implemented in literature [1]. Hierarchical clustering techniques are generally viewed as more reliable than other types of techniques [6]. The cause is that it is a bottom up procedure which primarily implies that each document is a cluster and thus merges the best similar clusters in an iterative procedure. The total number of iterations relies upon k, the number of required clusters. Due to its quadratic computational complexity, hierarchical clustering techniques are un-practical for large collection of documents. A number of techniques that attempt parallelism in hierarchical clustering techniques have been introduced in research[5-10]. A similar hierarchical clustering method is introduced in [3] which is applied as the clustering sub-routine for a parallel buckshot method. Instructive statistics are utilized to decrease the illustration of cluster centroids and thus decrease communication expense.
The random cluster centroid based technique is the most well-known summarization technique used to find the inter and intra document relationships in the large corpus. MEAD is a clustering technique based on the cluster-centroid approach for the multidocument summarization process. It is based on a phrase or sentence extraction. For each phrase or sentence in the documents, MEAD system calculates three characteristics and uses a cluster linear combination of the sentences and phrases. The three characteristics used are centroid score , overlap with the first sentence /phrase and positioned . For single document or group of phrase clusters it calculates the centroid based topic categories using tf-IDF type data.
The summarization approaches can be categorized into two types: unsupervised approaches, based upon the characteristics and heuristics resulting from the document text or sentence and supervised approaches, that depend upon machine learning techniques trained on preexisting summary documents, and. Supervised summarization approaches[12,13,14] handle the summarization activity as a two-class classification drawback at the phrase extraction stage, where the phrases and summary sentences are positive analytical samples while the phrases and non-summary sentences are negative analytical samples. It scores sentence clusters and phrases by merging sentence scores against text centroid , and tf-idf title overlap and text position value. The sentence is restricted by a user specified threshold and unused cluster sentences by checking cosine/Euclidean similarity towards earlier ones [11].
After symbolizing each sentence text by a vector of characteristics, the identification function can be trained in two different techniques[15]. The first is in a discriminating way with well known techniques such as SVM [14]. In [12], mathematical regression model,the use of genetic algorithm ,probabilistic neural network, Gaussian mixture model and back propagation system for text summarization process have been inspected.
This approach is a trainable Summarizer, which considers several characteristics, including positive keyword,sentence relative length, sentence position,sentence inclusion of named entities, sentence centrality,negative keyword, sentence resemblance to the title, sentence inclusion of numerical data, the bushy part of the sentence and combined similarity for each sentence to make summaries. As per Lefever et. al. [08], the challenging feature of this task is that it is in general not known previously specifically how many clusters to generate , thus the use of a Fuzzy Ants clustering method that does not depend on prior information of the number of clusters that need to be found in the documents. An analysis of benchmark data sets from SemEval's WePS2 or WePS1 competitions, shows that the resulting approach is successful with the agglomerative clustering method. Gorke et al. [17] implemented a procedure for grouping static or dynamic graphs. They use the minimum cut-trees to determine optimal clusters and present a technique to refine this data-structure when the graph nodes are updated. But, tree processing is an expensive operation and processing the initial tree is a global. Proposed tree has a high runtime and depends only on the size of the graph.

3.PROPOSED MODEL

Let N_p be the distributed peer to peer network with p nodes or peers. In any distributed p2p network, each peer can communicate with any other peer to exchange information like document sharing ,document clustering or document summarization etc.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

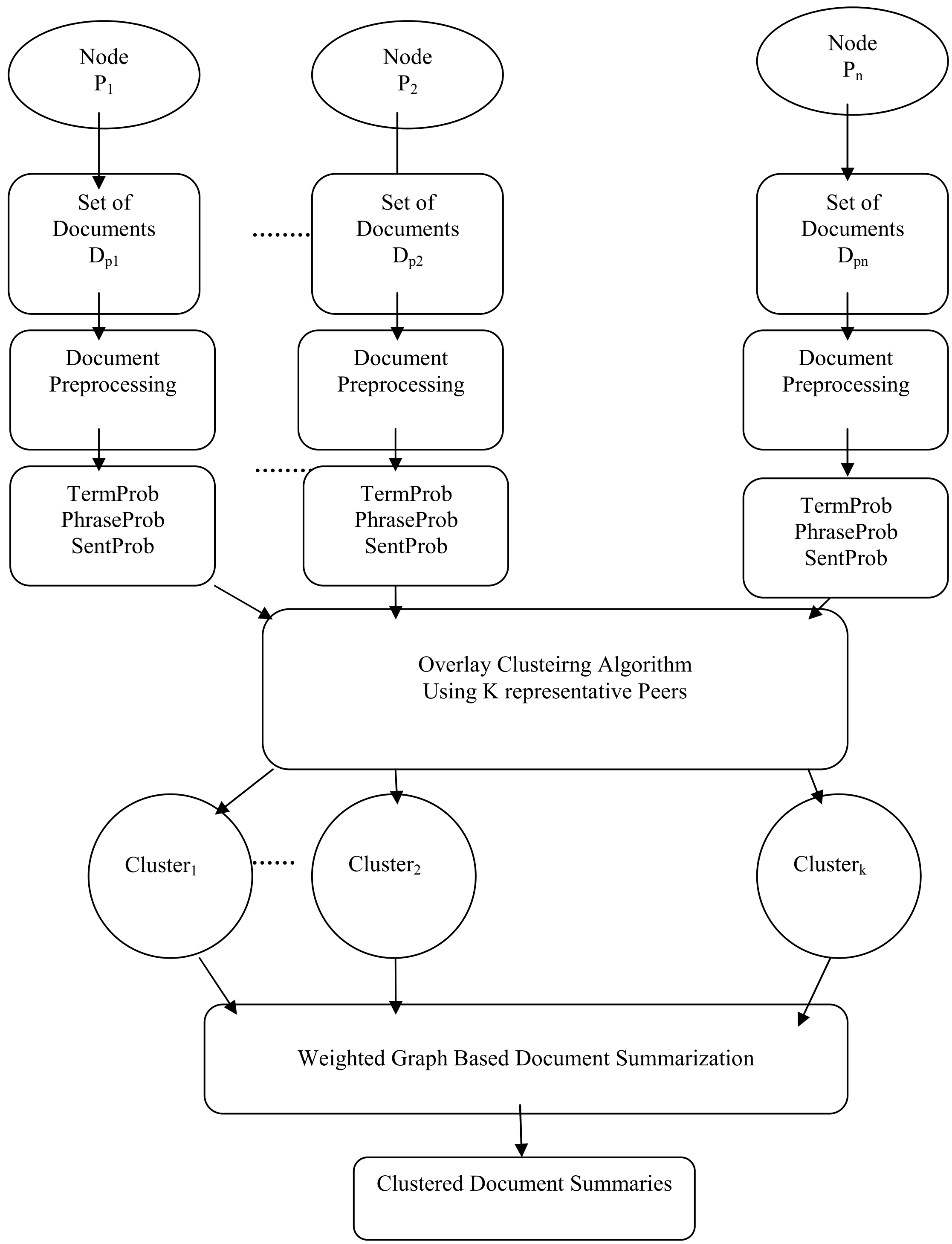


Figure 2. Workflow Of The Proposed Model

Journal of Theoretical and Applied Information Technology
20th August 2015. Vol.78. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1 Assumptions of the proposed model

The main aim of this work has been to optimize the graph methods for the improvement of the effectiveness of document clustering, by challenging some assumptions that implicitly characterize its application. Such assumptions relate to the static manner in which document clustering is typically performed, and include the static application of document clustering prior to querying, and the static calculation of inter-document p2p nodes. It does not require any previous knowledge, the approach extracts all required information from the domain.

3.2 Proposed Model

Each p2p network is partitioned into k overlays. Each overlay has a subset of nodes or peers to communicate with each other. In this approach, the global collection of documents is represented as D with n number of documents. Global document set D is partitioned into k overlay nodes. Proposed p2p document clustering and summarization process is described in figure 2 as shown above. In this approach, each overlay has set of peers or nodes associated with documents. In each node of the overlay network, set of documents are preprocessed to remove stopwords. After document pre-processing, document term probability, document phrase probability, document sentence probability are calculated to find the relevant clusters . Using the probabilistic document clustering algorithm, documents in the multiple overlay peers are grouped into k-representative clusters. After clustering operation is performed, all clustered documents with the peers are summarized to get final summaries.

3.2.1 Overlay probabilistic Document Clustering Algorithm:

Input:

O : set of overlays in the p2p network.
O_j : j^{th} overlay in the p2p network.
$P_{i,j}$: j^{th} peer or node in i^{th} overlay.
D_{ip} : set of documents in i^{th} overlay in p^{th} peer

Procedure:

Step 1: Initialize the p2p multiple overlay network.

For each $O_i, O_j \in O$ such that $O_i, O_j \neq \emptyset$ and $\mathbf{O}_i \cap \mathbf{O}_j \to \emptyset \ \forall i, j$

Step 2: Count number of documents in each overlay O.

Step 3: Select one random representative peer in each of the overlay.

Step 4: Let D_{ip} is the document in the i^{th} overlay of p^{th} peer.
For each node or peer p_i in O_j
Do
For each document D_{ip} in node or peer
Do
Tokenize(D_{pi}); // extract tokens, phrases and sentences
ExtractTokens();
ExtractPhrases();
ExtractSentences();
Done
Done

Step 5:
Calculate the Probabilistic Correlation Similarity between the peer and the representative peer of each overlay as
Let d_j and d_{ri} are the documents in the i^{th} peer and representative peer of k^{th} overlay.
α, β, γ be the document term correlation coefficient,phrase correlation coefficient and sentence correlation coefficient.
λ : cluster impact factor //user selected threshold
ε : Nodes coverage factor.
For each overlay O_k
Do
For each peer P_i
Do
Document term correlation coefficient α can be calculated as
// $\vec{\mathbf{d}}_{ip}$ representative of the document of the pth peer in i^{th} overlay.
For each term t in $\vec{\mathbf{d}}_{\mathbf{j}}$ of P_i
Do

Execute equation (1)

Add to Corr(P_i,t, $\vec{\mathbf{d}}_{\mathbf{j}}, \alpha_t$)
Done
Document Phrase correlation coefficient β can be calculated as
For each Phrase ph in $\vec{\mathbf{d}}_{\mathbf{j}}$ of P_i
Do

Execute equation (2)

Add to Corr(P_i, ph, $\vec{\mathbf{d}}_{\mathbf{j}}, \beta_{ph}$)
Done
Document Sentence correlation coefficient γ can be calculated as
For each Sentence s in $\vec{\mathbf{d}}_{\mathbf{j}}$ of P_i
Do

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Execute equation (3)

Add to Corr(P_i, s , $\vec{\mathbf{d}}_{\mathbf{j}}$, γ_s)

Done

The rank of the document can be givens as

$\sigma_{\mathbf{P}_i,\mathrm{r},\vec{\mathbf{d}}_{\mathbf{j}}} = (\alpha+\beta+\gamma)/1.5*(\max_i\{\alpha,\beta,\gamma\}+\max_{i+1}\{\alpha,\beta,\gamma\}$

If $\sigma_{\mathbf{P}_i,\mathrm{r},\vec{\mathbf{d}}_{\mathbf{j}}} > \lambda$ or $\sigma_{\mathbf{P}_i,\mathrm{r},\vec{\mathbf{d}}_{\mathbf{j}}} > (\lambda-\varepsilon)$

Then

AddCluster $\mathrm{C}_{ri}(\mathrm{P}_i, r, \sigma_{\mathbf{P}_i,\mathrm{r},\vec{\mathbf{d}}_{\mathbf{j}}})$ // r representative clusters in each overlay

End if

Done

Select new representative peer using sum of the documents rank of the peers from the initial representative peer.

$$r_k = \max\{\sum_{i=1}^{\mathbf{P}_i}\sum_{j=1}^{d}\sigma_{\mathbf{P}_i,\mathrm{r},\vec{\mathbf{d}}_{\mathbf{j}}}\}$$

ExtractTokens()

For each token t in D_{ip}

Do

Find the probability of t in D_{ip} as

Prob(t,D_{ip})=$|\, n(t,\mathbf{D}_{ip}) \cap n(t,\mathbf{D}_{jp})\,| \,/\, |\, n(t,\mathbf{D}_o)$

$\mathbf{D}_o$: Documents in the other peer of the current overlay

$\mathbf{D}_{jp}$: Documents in the j^{th} document of p^{th} peer of the current overlay

$\mathbf{D}_{ip}$: Document in the current peer of p^{th} peer of the current overlay

Done

ExtractPhrases()

For each Phrase ph in D_{ip}

Do

Find the probability of ph in D_{ip} as

Prob(ph,D_{ip})=

$|\, n(\mathrm{ph},\mathbf{D}_{ip}) \cap n(\mathrm{ph},\mathbf{D}_{jp})\,| \,/\, |\, n(\mathrm{ph},\mathbf{D}_o)$

ExtractSentences()

For each Sentences in D_{ip}

Do

Find the probability of s in D_{ip} as

Prob(s,D_{ip})=$|\, n(\mathrm{s},\mathbf{D}_{ip}) \cap n(\mathrm{s},\mathbf{D}_{jp})\,| \,/\, |\, n(\mathrm{s},\mathbf{D}_o)$

$\mathbf{D}_o$: Documents in the other peer of the current overlay

$\mathbf{D}_{jp}$: Documents in the j^{th} document of p^{th} peer of the current overlay

$\mathbf{D}_{ip}$: Document in the current peer of p^{th} peer of the current overlay

Done

$$\alpha_t = \frac{\{|\vec{\mathbf{d}}_{\mathbf{j}}|\sum_{i=1}^{|\vec{\mathbf{d}}_{\mathbf{j}}|}\mathrm{Prob}(t_j,\vec{\mathbf{d}}_{\mathbf{j}})*\mathrm{Prob}(t_r,\vec{\mathbf{d}}_{ir}) - \sum_{i=1}^{|\vec{\mathbf{d}}_{\mathbf{j}}|}\mathrm{Prob}(t_j,\vec{\mathbf{d}}_{\mathbf{j}})*\sum_{i=1}^{|\vec{\mathbf{d}}_r|}\mathrm{Prob}(t_r,\vec{\mathbf{d}}_r)\}}{\sqrt{(|\vec{\mathbf{d}}_{\mathbf{j}}|\sum_{i=0}^{|\vec{\mathbf{d}}_{\mathbf{j}}|}\mathrm{Prob}(t_j,\vec{\mathbf{d}}_{\mathbf{j}})^2)-(|\vec{\mathbf{d}}_r|\sum_{i=0}^{|\vec{\mathbf{d}}_r|}\mathrm{Prob}(t_r,\vec{\mathbf{d}}_r)^2)}} \quad \text{----(1)}$$

$$\beta_{ph} = \frac{\{|\vec{\mathbf{d}}_{\mathbf{j}}|\sum_{i=1}^{|\vec{\mathbf{d}}_{\mathbf{j}}|}\mathrm{Prob}(ph_j,\vec{\mathbf{d}}_{\mathbf{j}})*\mathrm{Prob}(ph_r,\vec{\mathbf{d}}_{ir}) - \sum_{i=1}^{|\vec{\mathbf{d}}_{\mathbf{j}}|}\mathrm{Prob}(ph_j,\vec{\mathbf{d}}_{\mathbf{j}})*\sum_{i=1}^{|\vec{\mathbf{d}}_r|}\mathrm{Prob}(ph_r,\vec{\mathbf{d}}_r)\}}{\sqrt{(|\vec{\mathbf{d}}_{\mathbf{j}}|\sum_{i=0}^{|\vec{\mathbf{d}}_{\mathbf{j}}|}\mathrm{Prob}(ph_j,\vec{\mathbf{d}}_{\mathbf{j}})^2)-(|\vec{\mathbf{d}}_r|\sum_{i=0}^{|\vec{\mathbf{d}}_r|}\mathrm{Prob}(ph_r,\vec{\mathbf{d}}_r)^2)}} \quad \text{-- (2)}$$

$$\gamma_s = \frac{\{|\vec{\mathbf{d}}_{\mathbf{j}}|\sum_{i=1}^{|\vec{\mathbf{d}}_{\mathbf{j}}|}\mathrm{Prob}(s_j,\vec{\mathbf{d}}_{\mathbf{j}})*\mathrm{Prob}(s_r,\vec{\mathbf{d}}_{ir}) - \sum_{i=1}^{|\vec{\mathbf{d}}_{\mathbf{j}}|}\mathrm{Prob}(s_j,\vec{\mathbf{d}}_{\mathbf{j}})*\sum_{i=1}^{|\vec{\mathbf{d}}_r|}\mathrm{Prob}(s_r,\vec{\mathbf{d}}_r)\}}{} \quad \text{.....(3)}$$

Journal of Theoretical and Applied Information Technology
20th August 2015. Vol.78. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\sqrt{(|\vec{\mathbf{d}}_{\mathbf{j}}|\sum_{i=0}^{|\vec{\mathbf{d}}_{\mathbf{j}}|}\mathrm{Prob}(s_j,\vec{\mathbf{d}}_{\mathbf{j}})^2)-(|\vec{\mathbf{d}}_r|\sum_{i=0}^{|\vec{\mathbf{d}}_r|}\mathrm{Prob}(s_r,\vec{\mathbf{d}}_r)^2)}$$

3.2.2 Graph Based Document Summarization Algorithm

Input :

Let G_k be the cumulative graph upto k documents.
Summ: Set of Document Summaries.
C : Set of Cluster Documents in all the overlays.
C_{ip} : Set of Subclusters in ith overlay of pth peer.
$\sigma_{C_i,\vec{\mathbf{d}}_{\mathbf{j}}}$: jth document score of ith cluster.
$\sigma_{C_i,\vec{\mathbf{d}}_{\mathbf{j}},\beta}$: Phrase score of jth document score of ith cluster.
$D_{C_i,\vec{\mathbf{d}}_{\mathbf{j}}}$: jth document of ith cluster.
θ : Candidate set factor //user defined value.

Procedure:

For each peer or node p_k in cluster i
Do
For each document in cluster i of p_k
$count$ =0;
If $\sigma_{C_i,\vec{\mathbf{d}}_{\mathbf{j}}} > 0$ and $count < \theta$
Then
CS=addCandidate($D_{C_i,\vec{\mathbf{d}}_{\mathbf{j}}}$, $\sigma_{C_i,\vec{\mathbf{d}}_{\mathbf{j}}}$);
end if
done
done
For each document $D_{C_i,\vec{\mathbf{d}}_{\mathbf{j}}}$ in candidate set CS.
For each Phrase ph_m in $D_{C_i,\vec{\mathbf{d}}_{\mathbf{j}}}$ // m phrases
Do
Terms[]=splitwords(ph_m);
v_1 =Terms[0];// initialize first term in vertex
If v_1 is not in Graph G
Add v_1 to Graph G.
Endif
For each term Terms[id] // id=2,3….len(terms)
Do
$v_{id} = Terms[\mathrm{id}]$
$v_{id-1} = Terms[\mathrm{id}-1]$
$e_{id} = (v_{id-1}, v_{id}, \sigma_{C_i,\vec{\mathbf{d}}_{\mathbf{j}},\alpha})$
If $v_{id} \notin$ G
Then
Add v_{id} to G
End if
If e_{id} $\in$ G then
For each cluster i
Get all document phrases ph_s from the cluster i which has score greater than e_{id}
Add a document phrase to Summ(ph_s);
Done
Else
Add edge e_{id} to Graph
End for
End for

4. EXPERIMENTAL RESULTS

Experimental results are performed on different datasets like 20NewsGroup, Reuters RC,Yahoo . Each data set has different types of categories along with documents.

Algorithm	Avg_Entropy	Avg_SeparationIndex	Accuracy
HP2PC	0.58	0.15	87
P2P K-means	0.76	0.23	69
MEAD	0.67	0.17	90
NeuralNetworks	0.76	0.29	88
GA_SVM	0.57	0.45	75
Proposed	**0.45**	**0.11**	**94**

Table 1: Distributeddataset

Table 1 describes the average entropy ,separation index and overall accuracy of distributed p2p network document clustering and summarization.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

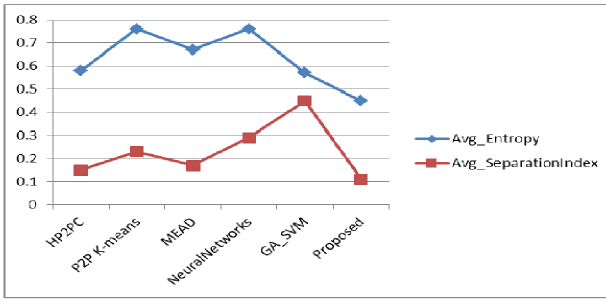


Figure 3 : Avg Separation Index Vs Average Entropy Of N- Peers

Algorithm	5 PEER-Accuracy	20 PEER-Accuracy	30-PEER Accuracy	40-PEER Accuracy
HP2PC	87	84	87	83
P2P K-means	69	79	78	74
MEAD	90	84	82	88
NeuralNetworks	88	76	72	79
GA_SVM	75	78	74	72
Proposed	**92**	**89**	**88**	**90**

Table 2: Summarization Accuracy In P2P Network

Table 2: Describes the overall clustering and summarization accuracy in the p2p network overlay. Proposed algorithm accuracy was compared with traditional robust approaches.

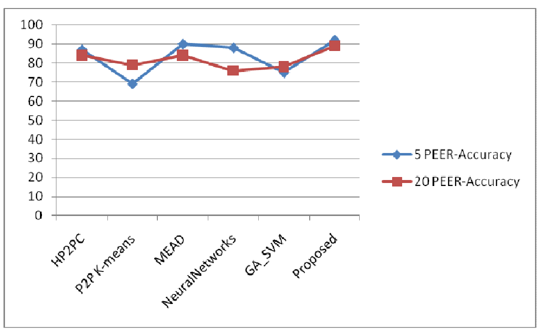


Figure 4 : 5 Peer Vs 20 Peer Accuracy Comparison

In the fig 4., p2p network with 5 peer and 20 peer are compared with traditional and proposed models. In the 5 peer networks, the proposed approach has high accuracy compared to existing methods. In the 20 peer network, proposed system runs in 89% accuracy.

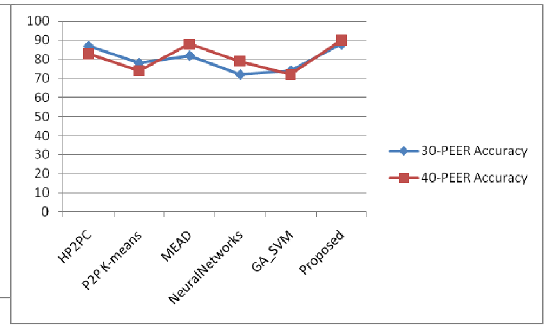


Figure 5 : 30 Peer Vs 40 Peer Accuracy Comparison

In the fig 5., p2p network with 30 peer and 40 peer are compared with traditional and proposed models. In the 30 peer networks, the proposed approach has high accuracy compared to existing methods. In the 40 peer network, proposed system runs in 90% accuracy.

Algorithm	5 PEER-RunTime(ms)	20 PEER-RunTime(ms)	30-PEER RunTime(ms)
HP2PC	3.45	17	22.67
P2P K-means	5.34	18.98	26.66
MEAD	3.76	16.87	24.76
Neural Networks	6.56	23.45	34.66
GA_SVM	6.34	24.86	44.66
Proposed	**2.65**	**16.78**	**22.45**

Table 3: 5,20,30 Peers Summarization Runtime

Table 3 describes the execution time of the summarization and clustering approaches in 5,20,30 peers overlay network.

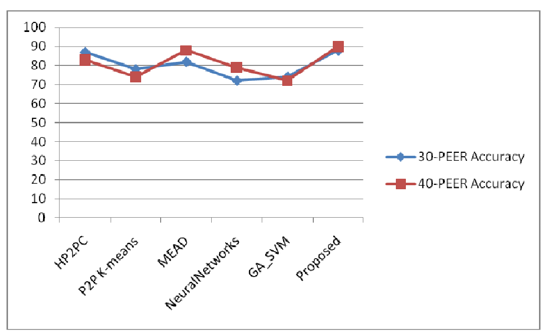


Figure 6 : Runtime Graphical Comparison

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. CONCLUSION

In this paper, a distributed p2p network was used to perform clustering based summarization process. Each overlay has dynamic overlays associated with peer nodes. Proposed clustering based summarization system works well on dynamic p2p networks Traditional cluster based summarization methods usually suffer with the computation speed, compression, peer selection and sentence clustering in order to generate high quality summaries. Traditional document clustering and summarization methods assume node adjacency and neighborhood information to build clusters and summaries. Proposed approach provides better solution to cluster different overlay networks by using probabilistic k representative clustering algorithm and forms efficient summaries using phrase rank based document summarization process. Experimental results give better performance in terms of execution time, entropy, similarity index are concerned.

REFERENCES

[1] M. J. McGill. And G. Salton ," Introduction to Modern Information Retrieval". Tata McGraw-Hill computer science series , 1983, New York.

[2] A. Wong, C. Yang and G. Salton,"A vector space model for automatic indexing". ACM Communications , 18(11):613–620, Nov1975.

[3] M. Littman and P. O. R. Allen, "An interface for navigating clustered document sets returned by queries". ACM Conference on Organizational Computing-Systems, 1993.

[4] O. Etzioni. And O. Zamir, " A dynamic clustering interface to Web search results". Computer Networks , 1999.

[5] M. Kamel and K. Hammouda," Distributed collaborative web document clustering using cluster key-phrase summaries". 2007. Information Fusion, Special Issue on Web Information Fusion, In Press.

[6] S. Sahni and S. Ranka, " Clustering methods on a Hypercube-Multicomputer", IEEE Transactions on Parallel and Distributed Computing, Vol. 2, No. 2, pp.129-137,2006.

[7] " Parallel algorithms for hierarchical clustering and applications to split decomposition and parity graph based recognition", Journal of Algorithms , E. Dahlhaus (2000), Vol. 36,pp. 205– 240.

[8] E. Jensen, S. Betzel, R. Cathey, O. Frieder & D. Grossman (2007), " Exploiting parallelism to support scalable hierarchical clustering", Society for Information-Science and Technology, Vol. 58, No. 8, pp. 127-1221.

[9] M. Kamel and K. Hammouda , "Collaborative Document Clustering methods" , In proceedings of 2006 SIAM Conference on Data Mining (SDM06), pp. 453-463, Bethesda.

[10] S. Rajesekanan, "Efficient Parallel Hierarchical-Clustering Methods", IEEE Transactions on Parallel and Distributed Systems, No. 6, Vol. 16, pp. 49-502.

[11] E. Hovy, D. Radev and K. McKeown, "Introduction to the special issue on summarization, Computational Linguistics", Vol. 28, No. 4, 2002.

[12] F. Ren and M. A. Fattah, "GA, MR, FFNN, PNN based models for automatic text-summarization", Computer Speech and Language.

[13] S. Ye, M.-Y. Kan, T.-S. Chua, and L. Qiu, "Document concept-lattice for text understanding and summarization", Vol. 43, Information Processing and Management, 2007, pp. 1643–1662.

[14] H-R. Ke, , I-H. Meng, J-Y. Yeh, W-P. Yang "Text summarization using a trainable summarizer and latent semantic analysis," Information Processing and Management, pp. 75–95, Vol. 41, No. 1, 2005.

[15] D. ShenH. Li, Q. Yang, J.-T.Sun, "Document summarization using conditional random-fields", (IJCAI'07), Hyderabad, India, pp. 862–287, January 6–12, 2007.

[16] Els Lefever, Veronique Hoste1, Martine De Cock Timur Fayruzov,"Fuzzy Ants-Clustering For Web People Search", Madrid, Spain, 3 April 20-24, 2009,