International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 30671-30686

http://www.ripublication.com

Medical Document Clustering And Summarization Approach In The Distributed P2P Overlays

A.Srinivasa Rao

Research Scholar
Dept of CSE JNTUK Kakinada,
A.P, India.

Dr.Ch.Divakar

Principal, Pydah college of engineering and technology,
Visakhapatnam, AP, India

Dr.A.Govardhan

Professor and Director, School of Information Technology,
JNTUH, Hyderabad, Telangana state, India.

ABSTRACT

User interest modeling is one of the essential services to the peer-to-peer document clustering and summarization systems. These systems can improve the efficiency of data sharing such as user recommended data clustering and summarization information. In the unstructured peer-to-peer overlay networks, large number of documents are clustered and randomly assigned to the distributed peers for summarization process. Complex user query may involve many peers and cause a high processing time for summarization. Unstructured peer-to-peer networks is basically suffering from high computational cost, high search cost and consumed more bandwidth and latency.Traditional document clustering and summarization process has high scalability and lack of peer cluster representation. In this work, we implemented a rich medical application as a solution for real-time document clustering and summarization using a graph based model. Experimental results show that proposed graph based document clustering and summarization model executes well against a large number of peers and documents.

Keywords: Document clustering, Summarization, Medline, P2P networks.

I.INTRODUCTION

The peer-to-peer system provides an environment for the sharing and management of data resources which are randomly distributed over the user's terminal. P2P systems are widely applied in distributed computing, file sharing and streaming services for the fault tolerance and scalability. Document clustering is a mechanism to assist keyword based similarity searching in peer-to-peer systems. Clustering and summarization in peer-to-peer overlay system is key area aimed to clear some benefits: One is increasing the accuracy of the document searching and clustering, and the other is to improve the quality of document summarizations. A lot of research has been introduced in the last few years for document clustering and retrieval techniques from the web.Due to the vast development in peer-to-peer frameworks and their usage in file sharing systems, document searching, document clustering and document summarization techniques.

Clustering algorithms are developed in machine learning domain to discover documents in the data, which generates a combination of different clusters according to similarity index. Documents with the same peer cluster are similar to each peer within the overlay and the documents in different peer clusters differ among them based on the user specified threshold. Overlay peer randomly selects clustered documents for summarization. Recently, three types of clustering algorithms have been used in peer-to-peer networks such as clustering by file type, clustering by locality awareness and clustering by file content. [1-4] proposed similarity based text clustering for adhoc mobile databases, aiming to generate an efficient process for data retrieval. They used the semantic concept based on data organization. Their method is based on local estimation of peers documents when they receive user query messages. Traditional class based semantic mechanism is used to cluster a large set of documents on a node into different classes. The design of effective information search and cluster methods is a centralized issue in these peer-to-peer networks. This is particularly useful for p2p applications that often require search based or similarity based document retrieval.All the documents in the peers are tagged with full semantic description along with correlation documents.

Document summarization is the method of producing a unique summary by minimizing the document length. Both the document clustering in the peers and document summarization algorithms used to retrieve meaningful patterns from the large collection of documents. These summarization methods[3-7] can be implemented on static peer nodes in the overlay networks to summarize each document cluster's information.Traditional summarization techniques select the relevant phrase or sentences from the peer documents to form a unique summary. So these techniques usually rank the phrase or sentences in the peer documents according to their predefined characteristics such as term-frequency or inverse term frequency. Almost all traditional document summarization techniques represent a document collection as the phrase or sentence term or phrase matrix in which each row denotes the phrase or sentence-id and corresponding column represents the term. The main issue with these methods are that they ignore the textual context information and assumes the phrase or sentences as independent to each peer documents.Automatic document summarization approaches can be partitioned into two groups : supervised

and unsupervised. The traditional information retrieval system was implemented in three cases: peer document initialization, document clustering and document summarization. In the initialization step, large number of documents are initialized to peer networks. In the document clustering step, each overlay peer is clustered according to the cosine similarity measure on the document set of the peer as shown in Fig1.

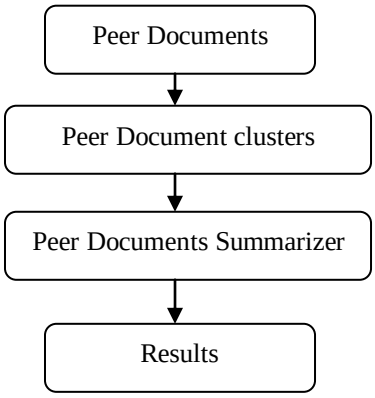


Fig 1: Traditional P2P system

A common technique to multidocument summarization is to extract a small number of phrases as a key sentence in the peer overlay. However, the traditional graph based models are applicable between the two sets of peer nodes and limited number of overlays. Therefore, it is not suitable to get the summarized information of all peers in the overlay network. Traditional LexRank and TextRank graph based models are application of the sentence summarization approach with the similarity index mechanism.

With the rapid development of online information, it becomes very difficult to find the essential information to the online users[2].MultiDocument summarization is the process of refining relevant information from a set of peer documents to produce a limited summarized information. In the paper[2] multidocument summarization method was implemented on the documents. In this system, each document is preprocessed and then document relevant features are extracted. After preprocessing, summarization approach applies to the document clusters to create a multidocument summary as shown in Fig2..

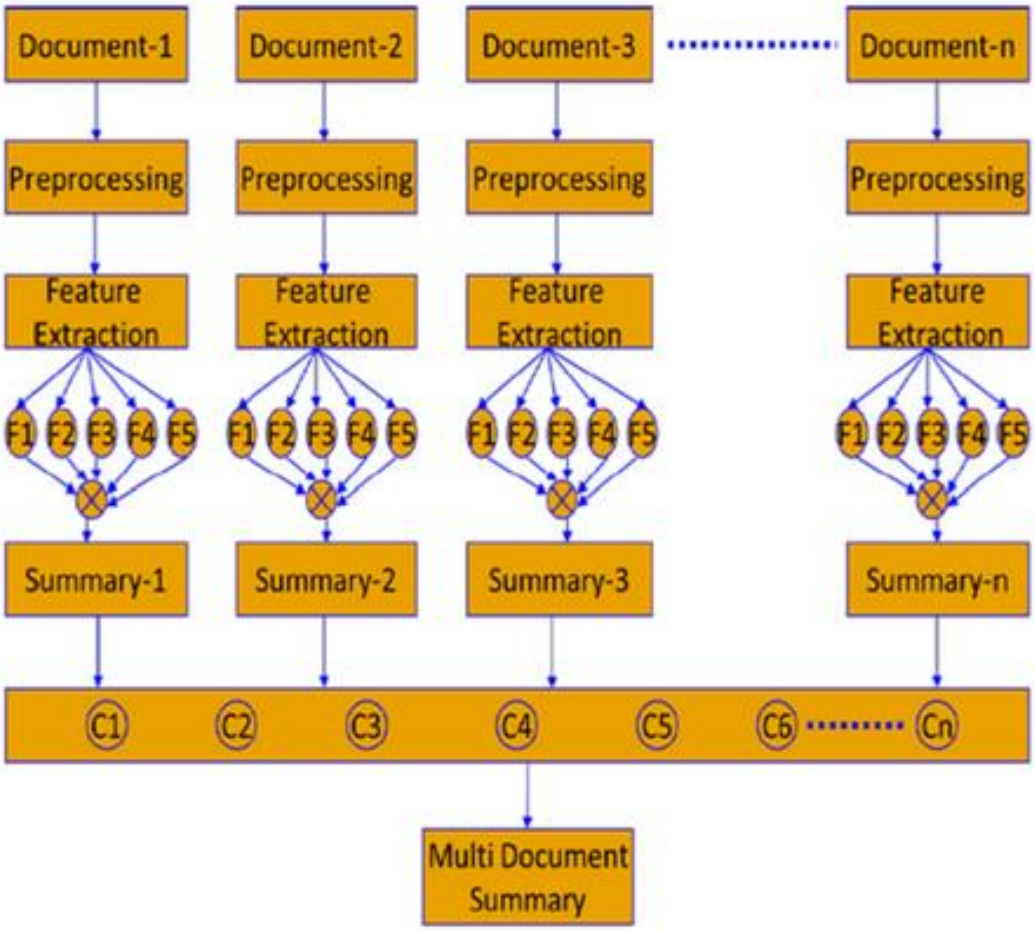


Fig 2: Multi-Document Summarization workflow

Model based clustering algorithms have been implemented on document clustering [4-9] where document clusters are represented as probabilistic methods that are conceptually separated from the data dimensions. Presently, graph based clustering models using statistical models also successfully applied to document clustering mechanism. The basic process of the p2p digraph clustering algorithm is to represent the documents as a directed graph in which each object is denoted as a graph node and the similarity weights on the edge denotes the document similarities between the peer documents. These graph models are optimized by using some predefined document measures on the directed graph.

Most of the present summarization system is focused on query based summarization. In the query based summarization, each phrase or sentence scores and term frequencies are used to estimate the topic's importance. The hierarchical LDA method was implemented in [10] as unsupervised method which is a generalization of LDA. Graph ranking based summarization algorithms have been implemented to construct a sentence model graph in which each node is a sentence in the overlay

documents. The traditional document summarization algorithms are not suitable due to following reasons.

1) Most of the algorithms work in a batch processing which causes inefficiency.
2) An iterative process will merge the each iteration summary into the existing summary, this type of solution would raise the duplicate issue.

Centroid based algorithms such as non-negative vector factorization, MEAD, Conditional random field based document summarization are introduced and the results are not satisfied due to the static initialization vector and the ranking approach is difficult to satisfy the summaries.

Factorization with the given bases(FGB) based document clustering cum summarization model was implemented in [3].This model minimizes the KL divergence between the peer documents and reconstructed terms.The document clusters are calculated by allocating each peer document to the phrase or sentence with the highest conditional probability value and the final summary is generated based on the high conditional probability of each document rank as shown in Fig3.

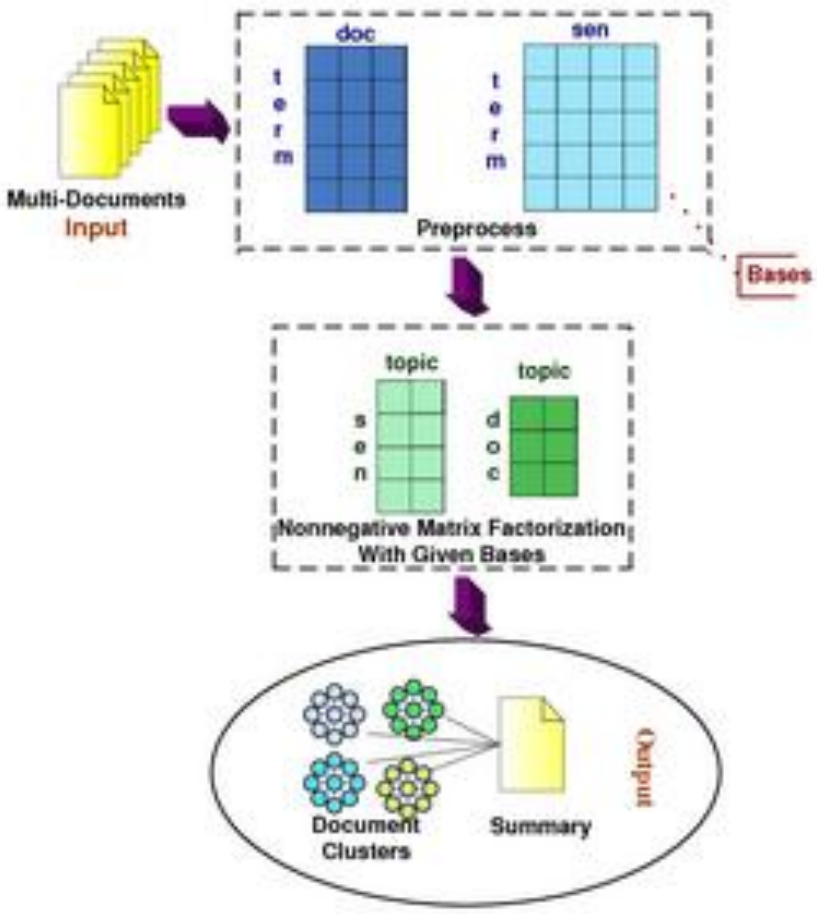


Fig 3: FGB document summarization process

Some techniques detect sentences of a document first and compute scores to the peer document sentences according to the relationship between the given topic and the sentences. The centroid based summarization model uses the peer center vectors to represent the topic's importance. Many features such as sentence position, centroid value and sentence overlap determine the sentence rank.The topic ranking model is a complex feature used to generate document topics with the large set of sentences. It consists of phrases or terms with bigrams and trigrams to the given topic. The score of the sentence is the sum of all the weights of term weights in the sentences. Extended kmeans clustering technique used to detect the typical clusters in the overlay clusters. Supervised learning models like artificial neural networks are used for document classification cum summarization process. The features obtained with this model not only determine sentence length, sentence position, most frequent terms and neighboring sentences, but also denotes complex characteristics like sentence scores and document ranking.

II. PROPOSED WORKFLOW

In this system, an improved document clustering and summarization approach is proposed using the graph based model was implemented. A medical application of the realtime P2P network was implemented on a large set of documents. In this application, each peer initializes set of documents for summarization. Graphical based P2P overlay application is used to visualize the k-clustered documents and then summary documents are created for each peer cluster.

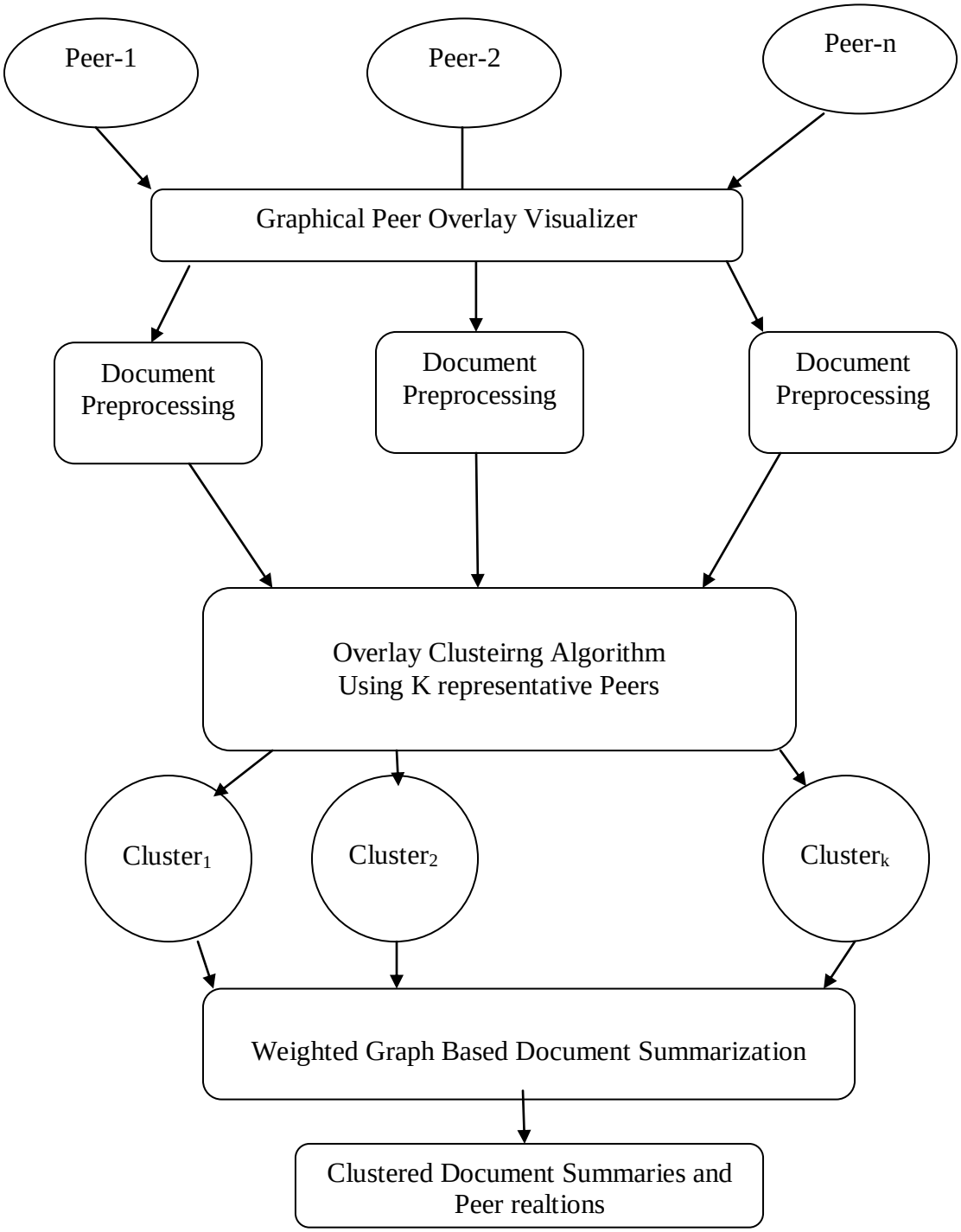


Figure 4. Graph based P2P Document Clustering based Summarizer

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 4 as shown above.

Graphical Visualizer Initialization:

Input:
Peers P,
Documents D,
Overlays O.

Output:
Graphical peers initializer.

Procedure:
Randomly initialize the overlay network with a set of peers.
For each peer in the network
Do
Randomly select the medical documents from the web.
For each category of documents
Do
Assign each medical document to the ith peer.
Done
Done
For each medical document in peer
Do
Generate_id(document)
Done
Create the initial graph nodes with the peer.
Execute document clustering algorithm.
Execute document summarization algorithm.

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 $O_i, O_j \neq \varnothing$ and $\mathbf{O}_i \cap \mathbf{O}_j \rightarrow \varnothing \ \ \forall i, j$
Step 2: Count number of documents in each overlay O.

Step 3: Select one random representative peer in each of the overlay as shown in Figure 4.
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:
Execute document clustering algorithm as specified in paper[4].

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 i^{th} overlay of p^{th} peer.
$\sigma_{C_i,\vec{\mathbf{d}}_j}$: j^{th} document score of i^{th} cluster.
$\sigma_{C_i,\vec{\mathbf{d}}_j,\beta}$: Phrase score of j^{th} document score of i^{th} cluster.
$D_{C_i,\vec{\mathbf{d}}_j}$: j^{th} document of i^{th} 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 $\sigma_{C_i,\vec{\mathbf{d}}_j} > 0$ and $count < \theta$
Then
CS=addCandidate($D_{C_i,\vec{\mathbf{d}}_j}$, $\sigma_{C_i,\vec{\mathbf{d}}_j}$);
end if
done
done
For each document $D_{C_i,\vec{\mathbf{d}}_j}$ in candidate set CS.

For each Phrase ph_m in $D_{C_i,\vec{\mathbf{d}}_j}$ // 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

$v_{id} = Terms[\text{id}]$

$v_{id-1} = Terms[\text{id}-1]$

$e_{id} = (v_{id-1}, v_{id}, \sigma_{C_i,\vec{\mathbf{d}}_j,\alpha})$

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 sum ();

Done

Else

Add edge e_{id} to Graph

End for

End for

III. Experimental Results

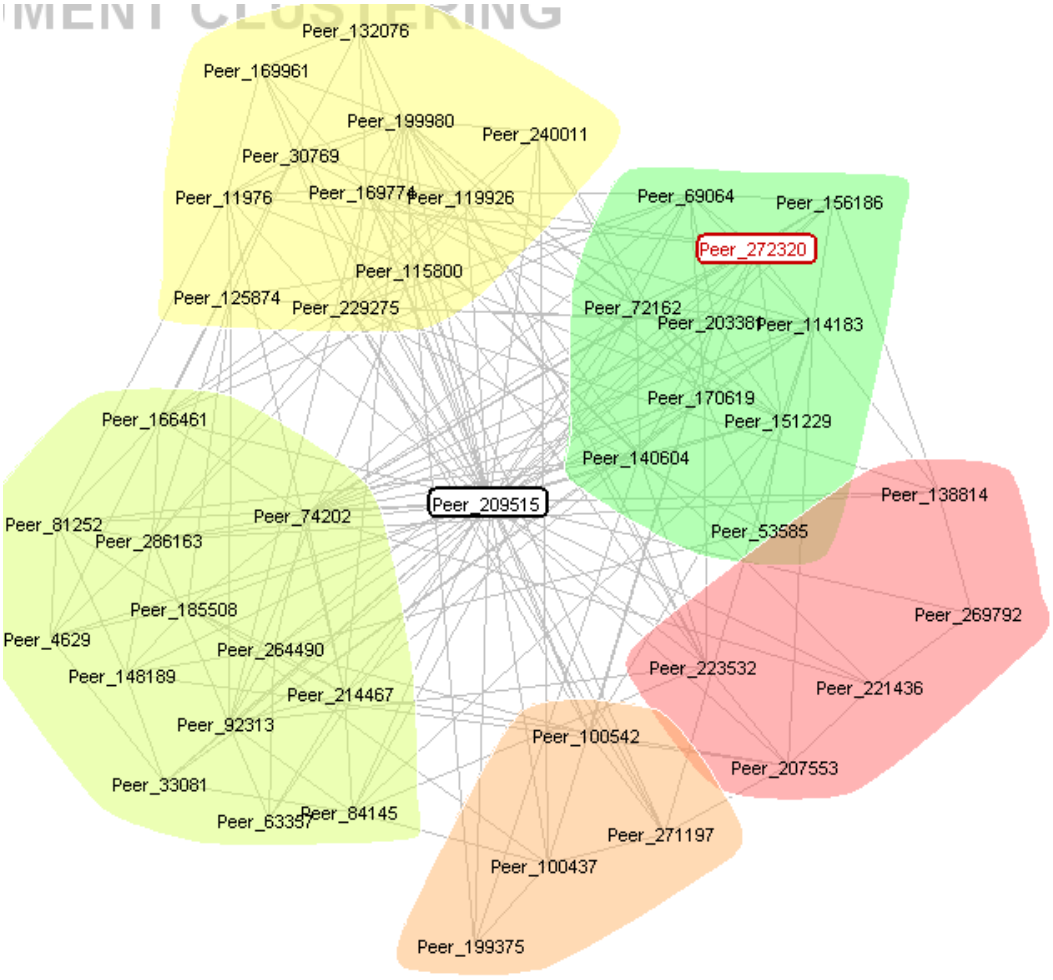
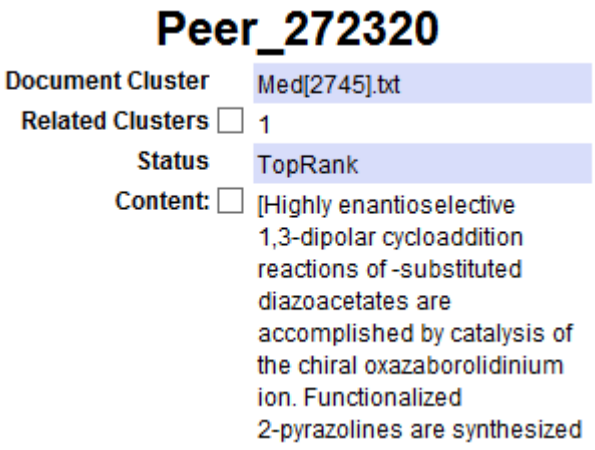


Fig 5: Peer Document Clustering



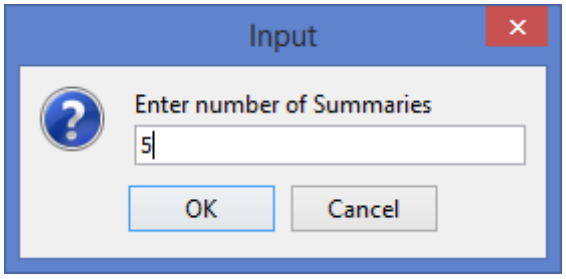


Fig 6: Number of summaries

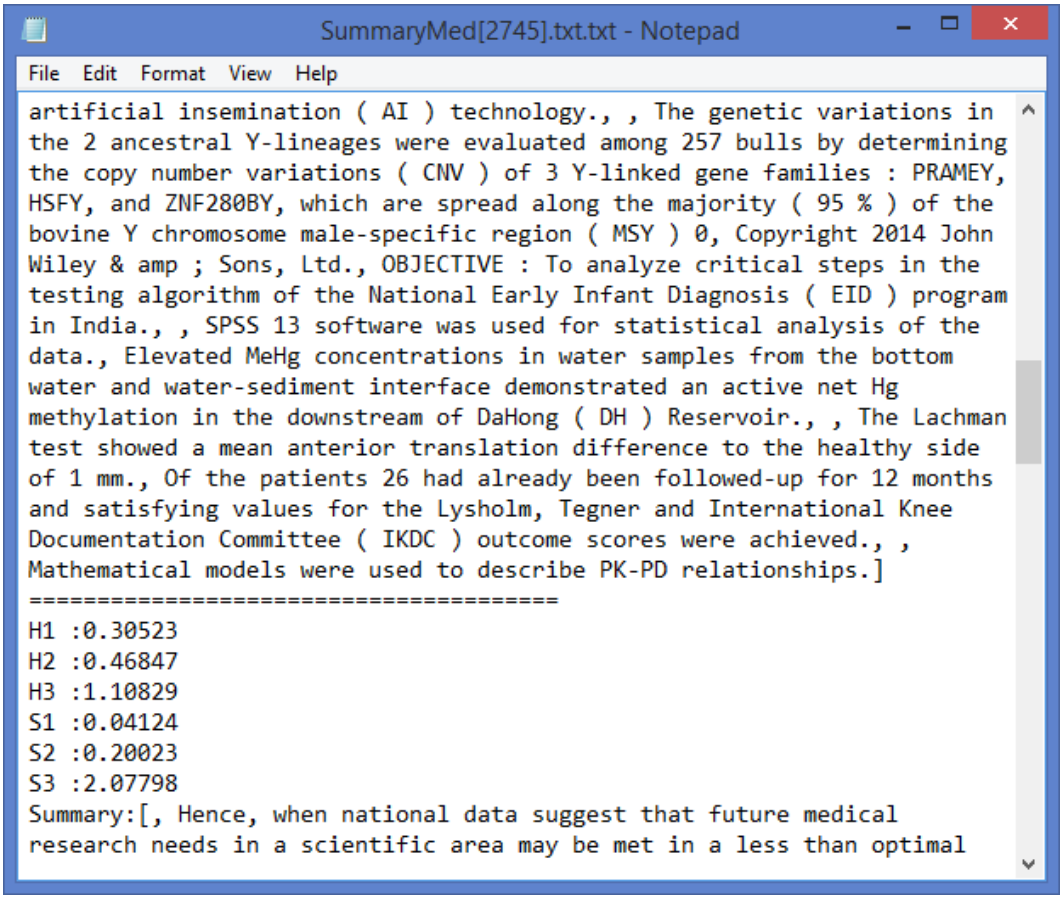


Fig 7: Final document summaries

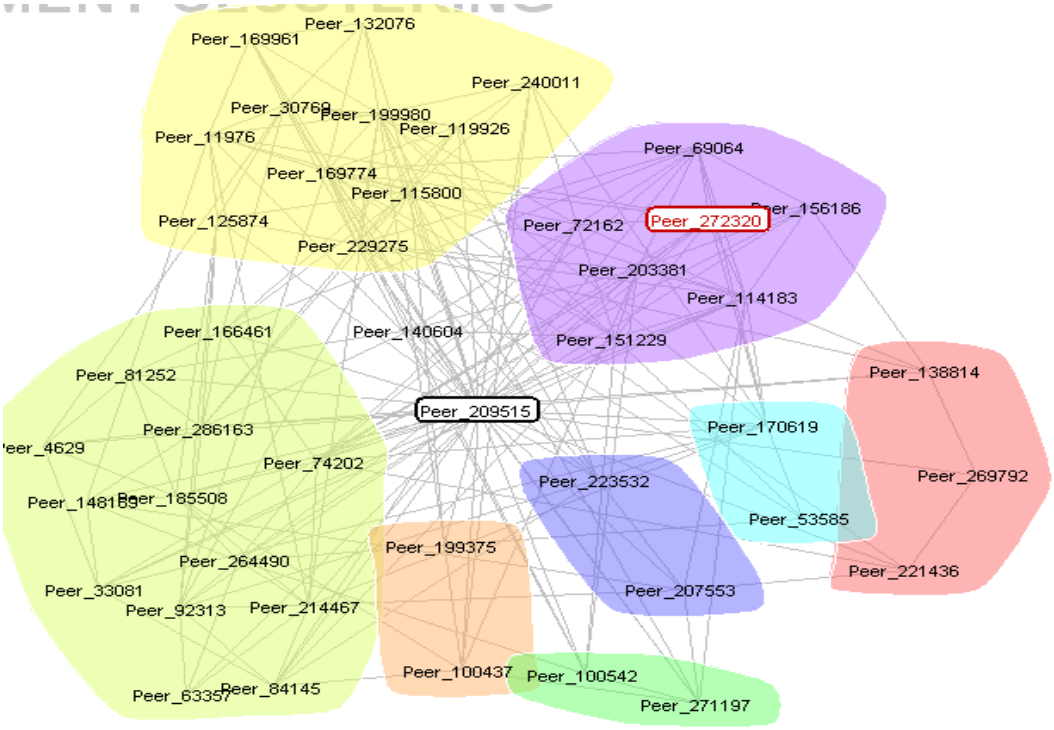


Fig 8: Dynamic P2P document clustering

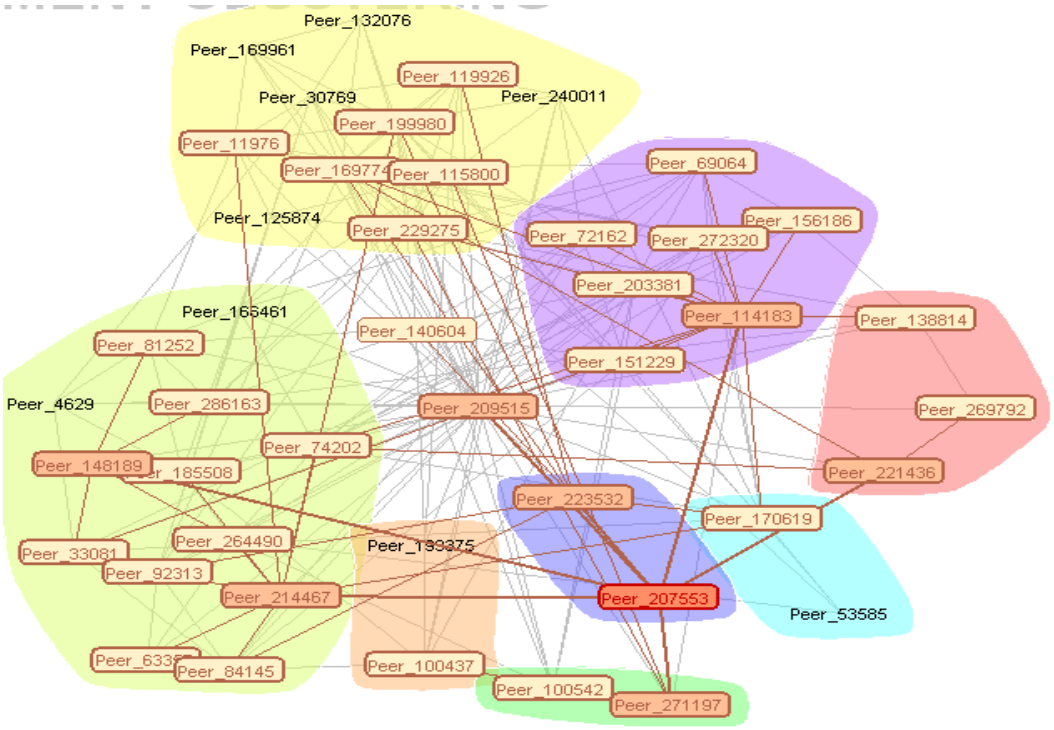


Fig 9: P2P overlay relations

Performance Analysis

Table 1: Runtime performance per peers

Peers	Documents	Summaries	RunTime(ms)
50	100	5	1987
75	150	5	2311
100	200	7	2534
125	250	7	2984

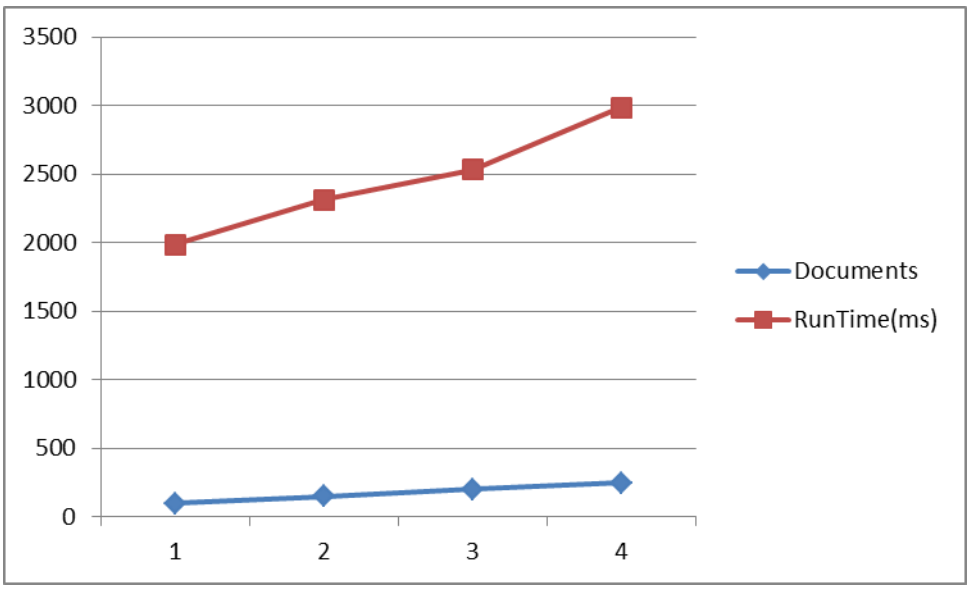


Fig 10: Graphical representation of documents vs runtime

Table 2: Cluster rate vs Accuracy

Clusters	Summaries	ClusterTime(sec)	Accuracy
4	5	0.5	0.956
5	5	0.67	0.965
6	7	0.72	0.959
7	7	0.98	0.969

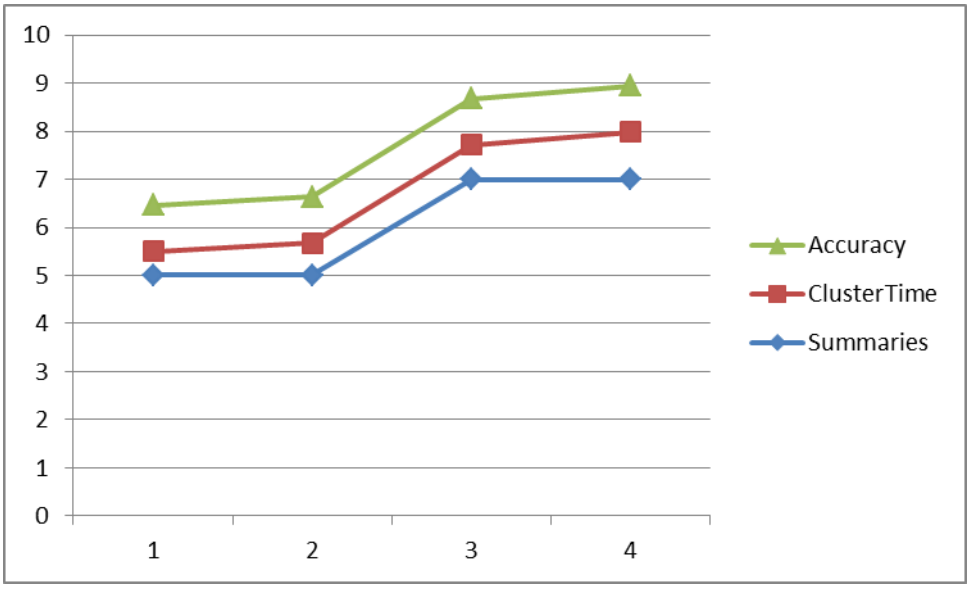


Fig 11: Graphical representation of cluster rate vs accuracy

Table 3: Medical Documents Clustering and Summarization Performance

Algorithm	5 PEER-Accuracy	20 PEER-Accuracy	30-PEER Accuracy	40-PEER Accuracy
HP2PC	83	81.65	83.23	82
P2P K-means	68.45	74	76.34	76.89
MEAD	87.89	86.34	85.67	87
NeuralNetworks	85.89	79.05	74.56	77.12
GA_SVM	74	75.78	76.45	79.45
Proposed	**91.56**	**88.12**	**88.97**	**90.29**

IV.CONCLUSION

In this research work, p2p application based clustering and summarization system was implemented on medical abstracts.Raw medical abstracts are large in number and it is very difficult to process manually. Proposed system was successfully implemented on medial abstracts for clustering and summarizing. This system takes less time to process clustering and to generate relevant summaries. Traditional document clustering and summarization process has high scalability and lack of peer cluster representation. In this work, we implemented a rich medical application as a solution for real-time document clustering and summarization using a graph based model. Experimental results show that proposed graph based document clustering and summarization model executes well against a large number of peers and documents.

REFERENCES

[1] J. da Silva, C. Giannella, R. Bhargava, H. Kargupta and M. Klusch, "Distributed Data Mining and Agents, " Eng. Applications of Artificial Intelligence, vol. 18, no. 7, pp. 791-807, 2005.

[2]. Multi-Document Summarization Using Sentence Clustering, Virendra kumar gupta, Samsung india software operation bangalgore, India.IEEE proceedings, 2012.

[3]. ”Integrating clustering and multi-document summarization to improve document understanding”, Dinding wang, Tao li, shenghuo zhu, Yun chi, Yihong Gong, ACM proceeding, 2008.

[4] A.Srinivasa Rao,Dr. Ch. Divakar, Dr.A.Govardhan,"Rank based document clustering and summarization approach in the distributed P2P network",JATIT,Vol 78, 2015.

[5] Sanghamitra Bandyopadhyay, Chris Giannella, Ujjwal Maulik, Hillol Kargupta, Kun Liu, and Souptik Datta. Clustering distributed data streams in peer-to-peer environments. Information Sciences, 176(14):1952–1985, 2006.

[6]. Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, Hai-Yan Liu, ”The heavy frequency vectorbased text clustering”, International Journal of Business Intelligence and Data Mining, 2005, Vol. 1, No.1 pp. 42 - 53.

[7] J. Branch, B. Szymanski, C. Gionnella, R. Wolff, and H. Kargupta. “In-network outlier detection in wireless sensor networks”. In ICDCS, 2006.

[8] S. Datta, C. Giannella, H. Kargupta, “Approximate Distributed K-Means Clustering over a Peer-To-Peer Network.” IEEE TKDE, 21(10), pp.1372-1388, 2009.

[9] S. Datta, K. Bhaduri, C. Giannella, R. Wolff, and H. Kargupta, “Distributed Data Mining in Peer-to-Peer Networks, ” IEEE Internet Computing, vol. 10, no. 4, pp. 18-26, 2006.

[10]. Hisham Al-Mubaid, Syed A. Umair, ”A New Text Categorization and Technique using distributional clustering and learning Logic”, IEEE Transactions on Knowledge and Data Engineering, 2006, Vol 18 Issue 9 pp 1156-1165.