Query-based Multi-Document Summarization by Clustering of Documents

Naveen Gopal K R
Dept. of Computer Science and Engineering
Amrita Vishwa Vidyapeetham
Amrita School of Engineering
Amritapuri, Kollam -690525
nvngkr@gmail.com

Prema Nedungadi
Amrita CREATE
Dept. of Computer Science and Engineering
Amrita Vishwa Vidyapeetham
Amrita School of Engineering
Amritapuri, Kollam -690525
ammasprema@gmail.com

ABSTRACT

Information Retrieval (IR) systems such as search engines retrieve a large set of documents, images and videos in response to a user query. Computational methods such as Automatic Text Summarization (ATS) reduce this information load enabling users to find information quickly without reading the original text. The challenges to ATS include both the time complexity and the accuracy of summarization. Our proposed Information Retrieval system consists of three different phases: Retrieval phase, Clustering phase and Summarization phase. In the Clustering phase, we extend the Potential-based Hierarchical Agglomerative (PHA) clustering method to a hybrid PHA-ClusteringGain-K-Means clustering approach. Our studies using the DUC 2002 dataset show an increase in both the efficiency and accuracy of clusters when compared to both the conventional Hierarchical Agglomerative Clustering (HAC) algorithm and PHA.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Query formulation*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms, Similarity measures*

General Terms

Algorithms, Performance, Experimentation

Keywords

Information Retrieval, Automatic Text Summarization, Hierarchical Agglomerative Clustering algorithms, Potential-based Hierarchical Agglomerative clustering, k-means, Clustering Gain

ICONIAAC '14, October 10 - 11 2014, Amritapuri, India

http://dx.doi.org/10.1145/2660859.2660972

1. INTRODUCTION

Automatic Text Summarization [12] reduces the volume of information by creating a summary from one or more text documents. The focus of the summary may be either generic, which captures the important semantic concepts of the documents, or query-based, which captures the sub-concepts of the user query and provides personalized and relevant abstracts based on the matching between input query and document collection. Currently, summarization has gained research interest due to the enormous information load generated particularly on the web including large text, audio, and video files etc. Text Summarization provides an overview of a large text allowing the user to understand, and reject or include the text without reading the whole text. Summarization can be useful in text classification, question answering, information retrieval etc. Search engines such as Google use summarization techniques for improving their search quality.

Generally, Automatic Text summarization methods can be divided into two categories: supervised and unsupervised methods. Summarization tasks may be extractive or abstractive. Extractive methods extract significant sentences from the given documents and generate a summary while abstractive methods may further modify the sentence structure.

In this work, we implement a more efficient and integrated Information Retrieval system [1] [2] [3] with three different phases by an extractive based query oriented multi-document summarization. The three phases include Retrieval phase, Clustering phase and Summarization phase. The methods used in these phases are all unsupervised methods and do not require any training data. In the Retrieval phase, a keyword matching links the user query and with the document collection using cosine similarity as the similarity measure. This means that the top-scored relevant documents are retrieved. In the Clustering phase, the retrieved documents are clustered into different topic groups based on the score obtained in the first phase. The actual summary is formed in the Summarization phase. The sentences in each of these clusters are ranked using the ranking method, TextRank. We use sentence level extraction approach for the summarization that extracts top ranked sentences from each of the clusters and form the summary.

The efficiency of the conventional Hierarchical Agglomerative Clustering (HAC) [13] method, Single Linkage clustering and the Potential-based Hierarchical Agglomerative (PHA) clustering is frequently reduced by the chaining problem and form heterogeneous clusters that may affect the performance of the retrieval system. Another drawback of HAC concerns the Convergence Criterion. Alternative methods also face problems. Choosing the optimum number of Clusters k restricts the efficiency of Partitioning-based approach such as k-means method. However, HAC methods and Partitioning-based clustering methods can be combined for efficient and accurate clustering. In this work, we extend PHA clustering method to a PHA-ClusteringGain-K-Means hybrid clustering approach, in which the PHA method and k-means method are combined by incorporating a clustering evaluation criterion concept called clustering gain into PHA for finding the optimum number of clusters automatically. This number is assigned as the value of k in k-means algorithm and forms homogeneous clusters that in turn eradicates the above disadvantages and improves the performance. The proposed hybrid clustering approach is more accurate and efficient than the conventional HAC method and the PHA method. The Clustering module in the proposed system is tested using Silhouette Coefficient method and the Summarization part is tested using ROUGE evaluation method. All sets of experiments are conducted using DUC 2002 dataset.

2. RELATED WORKS

There are many existing Information Retrieval systems with different phases such as Retrieval phase, Clustering phase and Summarization phase. The Clustering module in such systems clusters the retrieved documents into different topic groups and improves the retrieval system by providing organized and focused information. The Summarization module provides an abstract of large documents and allows users to find relevant information quickly without reading the whole text.

QCS [4], a system for querying, clustering and summarizing documents, is an Information Retrieval system that employs three phases Querying phase, Clustering phase and Summarization phase. In Querying phase, QCS retrieves a set of relevant document for a given input query using Latent Semantic Indexing (LSI). In Clustering phase, the retrieved documents are clustered into different topic clusters using generalized spherical k-means algorithm. In Summarization phase, a summary is created from each clusters using the methods Hidden Markov Model (HMM) and pivoted QR decomposition. However, k value used in the spherical k-means algorithm is given as user input.

Hierarchical Agglomerative Clustering (HAC) methods have been extensively applied in the field of Information Retrieval as the methods can improve the efficiency of the Information Retrieval systems. Anastasios Tombros, Robert Villa, and C.J. Van Rijsbergen [8] proposed a method for improving the effectiveness of retrieval systems that use Hierarchical Agglomerative clustering (HAC) methods for clustering the search results (query specific clustering).

However, the clustering methods used in these systems are affected by many disadvantages listed below. The proposed hybrid clustering method for clustering solves these disadvantages.

3. DISADVANTAGES CLUSTERING METHODS

Hierarchical Agglomerative Clustering (HAC) methods such as PHA clustering method, Single Linkage clustering method and Partitioning-based clustering method such as k-means method have many disadvantages as listed as follows.

3.1 Chaining effect

Chaining is a common problem in Single Linkage clustering and PHA method and it can be defined as the gradual growth of a single cluster as one data object with the elements added to that cluster at each iterative step of the algorithm. This leads to the formation of impractical heterogeneous clusters and may result in the unequal partitioning of data objects. In a clustering process many singleton clusters are formed as a result of chaining. Thus the output clusters cannot be properly define from the input data objects. Example for chaining is given below.

Let the data objects be *p1*, *p2*, *p3*, *p4*, *p5* with random coordinate values as *[4, 5, 6, 5]*, *[3, 4, 5, 8]*, *[1, 2, 4, 5]*, *[2, 3, 6, 7]*, *[4, 3, 5, 6]* respectively. Let the Hierarchical Agglomerative clustering be performed on these objects and the clusters at each iteration of the algorithm is shown below.

First Iteration - [[p1], [p3], [p5], [p4,p2]]
Second Iteration - [[p1], [p3], [p4, p2, p5]]
Third Iteration - [[p3], [p4, p2, p5, p1]]
Fourth Iteration - [[p4, p2, p5, p1, p3]]

In each iteration, the single cluster itself is growing gradually with the new data objects added to it in the successive iteration. We obtain heterogeneous partitioning of data objects at each iteration. In a document clustering process if a cluster is formed with large number of less scored documents, then the output summary may contain more information for a less significant topic. This affects the performance of the Information Retrieval system in a way that the system gives more significance for particular information that may have less significance in reality.

3.2 Convergence criterion for HAC

The HAC methods merge two most similar data objects into a single cluster hierarchically in each iteration as bottom-up manner in a hierarchical tree. It starts from the bottom level with N clusters and go each level up and ends in the top level with a single cluster. The stopping criterion is a problem in HAC that is in which level the iteration needs to be stopped.

3.3 Deciding the number of clusters

Deciding optimum number of clusters k is the major problem with Partitioning-based clustering approaches such as k-means method.

4. PROPOSED SYSTEM

The proposed system solves the above disadvantages and the procedure of our system is given in algorithm 1.

Each of the phases in the system is explained in detail as follows.

4.1 Retrieval phase

Retrieval phase is first phase in our proposed system. In this phase, an input query is matched with a set of doc-

Algorithm 1 PHA-ClusteringGain-K-Means Hybrid Clustering Method

Input: Given document collection and input query
Output: Summary from the set of clusters

1. Retrieval phase
   (a) Pre-processing the given document collection.
   (b) Pre-processing the input query.
   (c) Find the set of matching documents with the input query using cosine similarity.
2. Clustering phase
   (a) Perform Potential based Hierarchical Agglomerative clustering (PHA) method on the obtained set of matched documents with corresponding cosine similarity scores.
   (b) Compute the clustering gain at each iterative step of the algorithm.
   (c) Fix the no. of clusters when the clustering gain reaches its maximum value.
   (d) Take this number of clusters as the value for k in the k-means algorithm.
   (e) Perform the k-means method with the computed k value.
3. Summarization phase
   (a) Rank the sentences from each of the obtained document clusters using TextRank.
   (b) Form the summary with the extracted top ranked sentences.

uments in the collection and retrieves a set of documents relevant to the query. Cosine similarity is used as the similarity measure for finding the similarity between the query and set of documents. Before computing the cosine similarity, the given document collection and the input query have undergone pre-processing in order to represent them in vectors using vector space model and is explained as follows.

4.1.1 Document pre-processing

Document pre-processing consists of several steps as follows,

- Segmentation: The given document collection is represented as a set of individual documents. Each individual document is converted to their standardized format. Set of terms are extracted from the document set.
- Stopword Filtering: Stopwords are removed using the Stopword list built by Gerard Salton and Chris Buckley at Cornell University. This word list is 571 words in length. Set of unique terms are extracted.
- Word Stemming: Porter's stemming algorithm is used to perform word stemming. Stemming can reduce the memory space required to store the words and makes computation easier.
- Input Matrix Creation: The given input document collection is represented by $m * n$ term-document matrix A, where m is the number of rows and n is the number of columns. Each row represents a set of unique terms (words) and each column represents a set of documents. Each column A_j represents weighted term vector of document j. Each element A_{ij} in the term-document matrix is a weighted $tf - idf$ value of term t in document d.

$$A_{ij} = TF - IDF(t, d) \tag{1}$$

$TF - IDF$ can be defined as,

$$TF - IDF(t, d) = TF(t, d) * IDF(t) \tag{2}$$

$TF(t, d)$ or term frequency is the ratio between number of times term t appears in document d and total number of unique terms.

$$TF(t, d) = \frac{n(t, d)}{|T|} \tag{3}$$

where $|T|$ is total number of unique terms.

$IDF(t)$ or inverse document frequency is defined as,

$$IDF(t) = \log(\frac{|D|}{d_t}) \tag{4}$$

where $|D|$ is the total number of documents in the given input document set and d_t is the number of documents containing the term t.

4.1.2 Query pre-processing

Query pre-processing consists of several steps as follows.

- Tokenizing: Input Query is tokenized into individual set of words.
- Stopword Filtering: Stopwords are removed using the Stopword list built by Gerard Salton and Chris Buckley at Cornell University.
- Spell checking: Spell checking on individual query words is done using enchant package in Python.
- Word Stemming: Porter's stemming algorithm is used to perform word stemming.
- Query Vector Creation: A $tf - idf$ vector is created for the query similar to the document vector.

4.1.3 Querying the documents

The j^{th} column of term-document matrix A represents the j^{th} document. The j^{th} document have the cosine similarity score s_j that determines the relevance of the documents to an input query q and is compute as,

$$s_j = \frac{q \cdot A_j}{\|q\|\|A_j\|} \tag{5}$$

where $0 <= s_j <= 1$. The document with similarity score greater than a cutoff value is relevant to a particular query.

4.2 Clustering phase

The Clustering phase is the second phase in our system, in which we extend the PHA clustering method to PHA-ClusteringGain-K-Means hybrid clustering approach for efficient and accurate clustering. Clustering organizes the retrieved documents into different topic groups. The topic can be defined as the terms given in the input query. Clustering improves the retrieval system by providing organized

and focused information. In the Clustering phase, the retrieved documents are clustered based on their cosine similarity scores computed in the first phase. The proposed method is described as follows.

4.2.1 PHA-ClusteringGain-K-Means hybrid clustering approach

In PHA-ClusteringGain-K-Means hybrid clustering appro ach, Potential based Hierarchical Agglomerative clustering (PHA) and k-means are combined. A clustering evaluation criterion concept called clustering gain is incorporated into PHA for finding the optimum number of clusters automatically. The k-means method takes this optimum number as the value of k instead of giving it as user input. The clustering gain is computed at each iterative step of PHA clustering method. The iteration is converged when the value of clustering gain reaches at its maximum value and the number of clusters at that stage is taken as the optimal number of clusters. The PHA method is affected by chaining problem, where as the data objects are partitioned homogeneously in k-means algorithm. The k-means method usually produces spherical shaped or symmetrical shaped clusters. So the final clusters obtained from the proposed hybrid clustering approach are free from the chaining effect and thus the disadvantages of clustering are solved. The PHA method and clustering gain computation are explained in detail as follows.

4.2.1.1 Potential-based Hierarchical Agglomerative Clustering (PHA).

Potential-based Hierarchical Agglomerative clustering [5] is a Hierarchical Agglomerative clustering method in which data objects are clustered based on a hypothetical potential field existing between all the data objects in the euclidean space. PHA considers both global data distribution information by incorporating potential field and local data distribution information by including distance matrix in the algorithm while conventional Hierarchical Agglomerative clustering algorithms such as Single Linkage clustering, Complete Linkage clustering only consider local data distribution information. In this work, cosine similarity scores of the retrieved documents computed in the first phase is taken as the data objects and euclidean distance is taken as the distance measure.

In the PHA method, the potential at a data object i from another data object j is defined as,

$$\phi_{ij}(r_{ij}) = \begin{cases} -\frac{1}{r_{ij}} & \text{if } r_{ij} >= \delta \\ -\frac{1}{\delta} & \text{if } r_{ij} < \delta \end{cases} \quad (6)$$

where $r_{i,j}$ is the euclidean distance between i and j and the parameter δ is for avoiding singularity when $r_{i,j}$ becomes zero. The δ value is compute as,

$$\delta = mean(MinD_i)/S \quad (7)$$

$$MinD_i = min_{r_{ij} \neq 0, j=1...N}(r_{ij}) \quad (8)$$

where $MinD_i$ is the minimum distance from data object i to all the other objects, and S is scale factor taken as 10. Total potential at a data object i is the sum of potential from all other objects and is computed as,

$$\phi_i = \sum_{j=1...N} \phi_{ij}(r_{ij}) \quad (9)$$

where N is the number of the data objects.

The procedure of PHA method is as follows.
Step 1: Let the data objects be *p1*, *p2*, *p3*, *p4*, *p5* and *p6*, representing cosine similarity scores of retrieved document, in euclidean space as shown in figure 1.

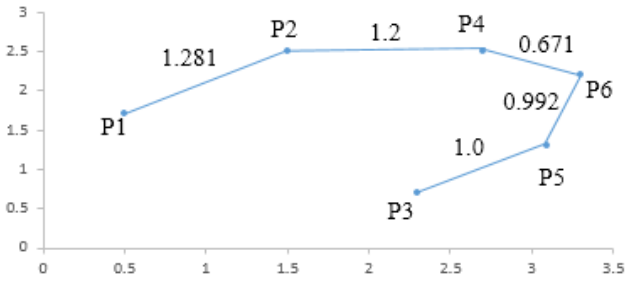

Figure 1: Data objects in the Euclidean space

Step 2: Compute the total potential at each object using equation (9) and sort these values in ascending order.

Step 3: Let the ascending order of potential be [*p4*, *p6*, *p5*, *p2*, *p3*, *p1*], which means the data object *p4* has the lowest potential become the root of the *Weighted_Edge_Tree* and the data object *p1* has the highest potential. Then, the method tries to find the parent of each data object. The parent of a given data object is defined as the nearest visited data object to it.

Step 4: The parent of second object *p6* is set as *p4* because *p4* is the nearest object visited so far.

Step 5: Parent of third object *p5* is taken as *p6* because *p5* is nearest to *p6* than *p4*.

Step 6: Similarly find parent of *p2* and so on up to *p1*.

Step 7: The *Weighted_Edge_Tree* is drawn as shown in figure 2. The weight of the edge between a data object and its parent is defined as the euclidean distance between them.

Step 8: The data objects are sorted based on the edge

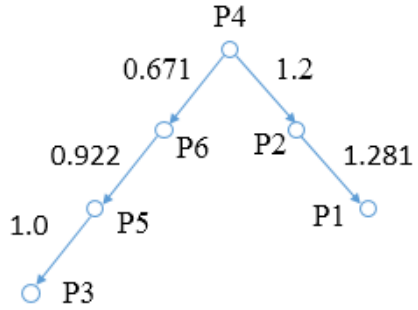

Figure 2: Weighted_Edge_Tree

weight between the data object and its parent as shown in the tree and let the order be [*p6*, *p5*, *p3*, *p2*, *p1*, *p4*]. This means that the distance between *p6* and its parent *p4* is the smallest.

Step 9: The first data object *p6* from the above queue is merged with its parent *p4* and form cluster *(p6, p4)*. The second data object *p5* is merged with the cluster *(p6, p4)*

because parent of *p5* is *p6*. The rest of the objects up to *p4* in the queue get merged with their respective parent.

Step 10: The given data objects are clustered hierarchically in each iteration and the clustering process is shown by a dendrogram in figure 3.

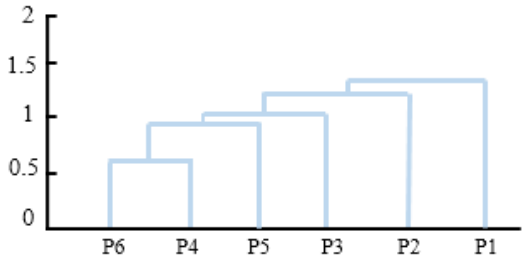

Figure 3: Dendrogram

The set of retrieved documents are clustered using PHA method into different topic groups. The PHA has the time complexity of $O(n^2)$ and is more efficient than conventional HAC method such as Single Linkage clustering with time complexity $O(n^3)$. The clustering gain computation is explained as follows.

4.2.1.2 Clustering Gain.

Clustering gain [7] can be regarded as a computationally efficient evaluation criterion for finding an optimal cluster configuration and there by finding an optimal number of clusters required in the clustering process. The value of clustering gain is always greater than or equal to zero. The optimal number of clusters can be computed at the maximum value of clustering gain. Complete clustering is needed to be done in order to determine the maximum value of clustering gain, as the value of clustering gain varies at each iteration of the clustering process from zero at initial and final stages of the clustering to maximum about at middle stage of the clustering. The clustering gain is computed at each iterative step of the PHA clustering method. The PHA method is converged at the maximum value of cluster gain. The no. of clusters obtained at that stage is taken as the value for k in k-means algorithm.

The Clustering gain, Δ_j computed for the cluster C_j in a particular stage of a clustering process can be defined as difference between the decreased inter-cluster error sum γ_j compared to the previous stage and increased intra-cluster error sum λ_j compared to the previous stage. This can be written as,

$$\Delta_j = \gamma_j - \lambda_j \tag{10}$$

where γ_j is the decreased inter-cluster error sum and λ_j is the increased intra-cluster error sum for the cluster C_j.

The intra-cluster error sum (within-group error sum) Λ is the sum of distances for all clusters such that the distance is the sum of squared Euclidean distances from the centroid of a cluster to every data objects in that cluster and is given by,

$$\Lambda = \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||p_i^{(j)} - p_0^{(j)}||_2^2 \tag{11}$$

where k is number of clusters, n_j is the number of instances in the j^{th} cluster, $p_i^{(j)}$ is the i^{th} instance in the j^{th} cluster and $p_0^{(j)}$ is the centroid point. The intra-cluster error sum is increased during a clustering process from zero at the initial stage and maximum at the final stage.

The inter-cluster error sum Γ is the sum of squared Euclidean distances from the centroid of a cluster to the global centroid for all clusters and is given by,

$$\Gamma = \sum_{j=1}^{k} ||p_0^{(j)} - p_0||_2^2 \tag{12}$$

where p_0 (global centroid) is the centroid of all data objects and is given by,

$$p_0 = \frac{1}{n} \sum_{i=1}^{n} p_i \tag{13}$$

4.2.1.3 K-Means clustering method.

The k-means [9] method is performed with the computed k value as input. The k-means method produces k no. of homogeneous clusters. Thus the output clusters obtained from the proposed hybrid clustering approach is free from chaining effect.

The time complexity of the proposed hybrid clustering approach, the sum of time taken by both PHA and k-means methods, is $O(n^2) + O(nidk)$, which is less than that of Single Linkage clustering with time complexity $O(n^3)$, where n is no. of instance in d dimension, i is the no. of instances and k is the no. of clusters to be obtained.

4.3 Summarization phase

The Summarization phase is the last phase in our proposed system, in which the actual summarization process is performed by ranking individual sentences from the obtained document clusters. Sentence level extraction approach is used for creating the summary. The summarization is done by forming a single summary from each of the clusters of documents by extracting top ranked sentences and presents a set of summary as the output of the system. The ranking method, TextRank is used for ranking the sentences.

4.3.1 TextRank

TextRank [6] is a graph based method for ranking sentences. These graphs are built from natural language texts with sentences in the texts form the vertices of the graph. The Edge e_{ij} between two vertices v_i and v_j represents the similarity between the i^{th} sentence and j^{th} sentence and the edge weight w_{ij} represents the value of similarity between them. In this work, the edit distance between two sentences is taken as the similarity value between them. In TextRank, PageRank method is employed for ranking purposes. The weighted PageRank method is called with the constructed graph as input. The weighted PageRank PR^W of vertex v_i is computed as,

$$PR^W(v_i) = (1-d) + d * \sum_{v_j \in In(v_i)} w_{ji} \frac{PR^W(v_j)}{\sum_{v_k \in Out(v_j)} w_{kj}} \tag{14}$$

where $PR^W(v_j)$ is weighted PageRank of v_j and d is a parameter having value between 0 and 1, usually set at 0.85. $In(v_i)$ is the set of vertices that points to v_i (predecessors)

and $Out(v_j)$ is the set of vertices that v_j points to (successors). Also w_{ji} is the edge weight between vertices j and i and w_{kj} is the edge weight between vertices k and j.

The PageRank algorithm starts with arbitrary values assigned to each vertex in the graph. The computation proceeds until a given threshold value is achieved. Importance of each vertex in the graph is indicated by a final value associated with them and the corresponding sentences are ranked based on these scores. Top ranked sentences are obtained by sorting the sentences in the decreasing order of their score.

5. EXPERIMENTATION AND RESULT

5.1 Dataset

In this work, DUC 2002 [10] dataset is used for the experimentation. The test data contains 59 document sets with each set containing 5 to 15 documents. For the experimentation purpose, document sets describing about natural disaster events such as hurricane, earthquake, flood, drought and volcano erupt are collected from the entire collection.

The performance comparison of our proposed hybrid clustering method with PHA clustering method and Single Linkage clustering algorithms is done using silhouette coefficient method. The silhouette coefficient obtained for the PHA clustering, the Single Linkage HAC and our proposed hybrid clustering method for different query is shown in table 1, 2, 3 respectively. The first column, no. of topics denotes the number of terms contained in a particular input query and the second column, number of clusters denotes the optimal number of clusters computed from the cluster gain for that query.

Table 1: Silhouette Coefficient values for PHA clustering

No. of Topics	No. of Clusters	Silhouette Coefficient
2	2	0.5691
3	3	0.6652
4	5	0.3505
5	6	0.5714

Table 2: Silhouette Coefficient values for Single Linkage HAC

No. of Topics	No. of Clusters	Silhouette Coefficient
2	3	0.2555
3	3	0.6652
4	5	0.3505
5	6	0.5714

Table 3: Silhouette Coefficient values for PHA-ClusteringGain-K-Means hybrid clustering method

No. of Topics	No. of Clusters	Silhouette Coefficient
2	2	0.5864
3	3	0.6652
4	5	0.5677
5	6	0.6049

It is clear from the above tables that our proposed hybrid clustering method obtains better values for silhouette coefficient than PHA clustering and Single Linkage HAC. So we can conclude that our method produced better clusters than that produced by PHA and Single Linkage clustering.

The time taken for execution by our proposed hybrid clustering method is compared with the Single Linkage HAC method for an input query *hurricane earthquake* and is given by a graph shown in figure 4.

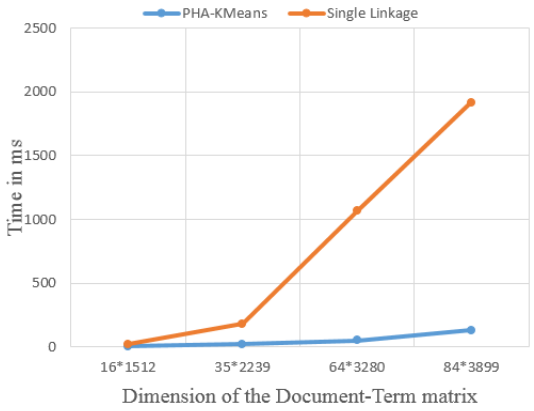

Figure 4: Time taken by PHA-ClusteringGain-K-Means hybrid clustering method and Single Linkage HAC

The X-axis denotes different dimensions of the input docum ent-term matrix and Y-axis denotes the time taken for execution in millisecond. It is clear from the graph that the time taken by our proposed method is significantly less than that of Single Linkage clustering under all dimensions of input document-term matrix. The obtained results agree with their theoretical time complexities

The clustering gain graph for input query *hurricane earthquake* is shown in figure 5. The X-axis denotes each iterative step of the propose hybrid clustering method and Y-axis denotes the clustering gain obtained at each iteration.

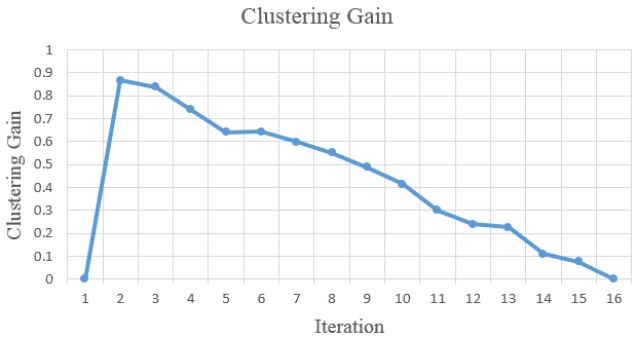

Figure 5: Clustering Gain

The chaining effect shown by the PHA clustering method is depicted by a graph in figure 6. Each point in the X-axis corresponds to the Silhouette Coefficient value of each data object (document) and the Y-axis corresponds to cluster number. Each cluster is colored with unique color. The PHA clustering is affected with chaining problem as it produce two clusters, as shown by blue and red color respec-

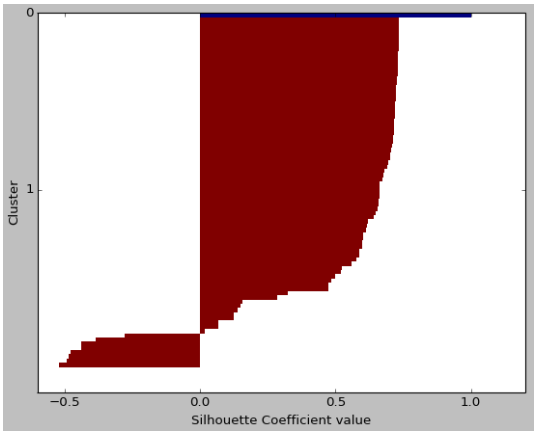

Figure 6: Chaining effect shown by PHA clustering

tively, with zeroth cluster contains only one data object (document) and first cluster contains rest of the data objects. This problem is solved by the proposed hybrid clustering approach as the method produced homogeneous clusters as shown in figure 7.

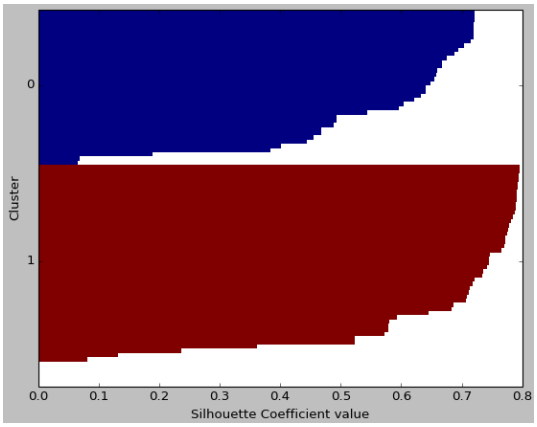

Figure 7: Homogeneous clusters obtained by PHA-ClusteringGain-K-Means hybrid clustering method

5.2 Summary evaluation measure

ROUGE [11] [14], a software package for automated evaluation of summaries, is used for evaluating the summary produced by our proposed hybrid clustering approach. The ROUGE evaluation is conducted for the input query *hurricane earthquake*. The query produces two clusters, with first cluster corresponds to the topic *hurricane* and second cluster corresponds to the topic *earthquake*. The Precision, Recall and F-Score for ROUGE-1, ROUGE-2, ROUGE-L is computed and given in table 4.

Table 4: ROUGE Evaluation

Rouge/Parameter	ROUGE-1	ROUGE-2	ROUGE-L
Precision	0.5000	0.3092	0.5000
Recall	0.4604	0.2846	0.4604
F-Score	0.4794	0.2964	0.4794

6. CONCLUSION

We implemented a more efficient and integrated Information Retrieval system with three different phases: Retrieval phase, Clustering phase and Summarization phase. In the Clustering phase, we extended the PHA clustering method to PHA-ClusteringGain-KMeans hybrid clustering approach by combing PHA method and k-means method by using the clustering gain evaluation criteria to determine the ideal k value for k-means. The DUC 2002 dataset is used for conducting the experimentation. Our results show that the proposed PHA-ClusteringGain-KMeans hybrid clustering approach is more efficient and accurate than conventional Hierarchical Agglomerative Clustering (HAC) algorithm and PHA.

7. ACKNOWLEDGMENTS

This work derives direction and inspiration from the Chancellor of Amrita University, Sri Mata Amritanandamayi Devi. We thank Dr. M. Ramachandra Kaimal, Head of Computer Science Department, Amrita University for his valuable feedback.

8. REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008, Introduction to Information Retrieval, *Cambridge University Press, New York, NY, USA*.

[2] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999, Modern Information Retrieval, *Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA*.

[3] David A. Grossman and Ophir Frieder. 1998, Information Retrieval: Algorithms and Heuristics, *Kluwer Academic Publishers, Norwell, MA, USA*.

[4] Daniel M. Dunlavy, Dianne P. O Leary, John M. Conroy, Judith D. Schlesinger, QCS: A system for querying, clustering and summarizing documents, *Information Processing and Management, ScienceDirect*, 2007. Volume 43, Issue 6, November 2007, Pages 1588–1605.

[5] Yonggang Lu, Yi Wan, PHA: A fast potential-based hierarchical agglomerative clustering method, *Pattern Recognition, ScienceDirect*, 2013, Volume 46, Issue 5, May 2013, Pages 1227–1239.

[6] Rada Mihalcea. 2004, Graph-based ranking algorithms for sentence extraction, applied to text summarization, *In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (ACLdemo '04), Association for Computational Linguistics, Stroudsburg, PA, USA, Article 20.*

[7] Yunjae Jung, Haesun Park, A Decision Criteria for the Optimal Number of Clusters in Hierarchical Clustering, 2002, *Kluwer Academic Publishers.*

[8] Anastasios Tombros, Robert Villa, & C. J. Van Rijsbergen, The effectiveness of query-specific hierarchic clustering in information retrieval, *Information Processing and Management, ScienceDirect*, 2002, Volume 38, Issue 4, July 2002, Pages 559–582.

[9] Velmurugan T., Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data, *Applied Soft Computing, ScienceDirect*, 2014, Volume 19, June 2014, Pages 134–146.

[10] DUC, 2002, Document Understanding Conference (DUC), 2002, http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html.

[11] Lin, Chin-Yew. 2004a, ROUGE: a Package for Automatic Evaluation of Summaries, *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain*, July 25 - 26, 2004.

[12] Ju-Hong Lee, Sun Park, Chan-Min Ahn, & Daeho Kim, Automatic generic document summarization based on non-negative matrix factorization, *Information Processing and Management, ScienceDirect*, 2009, Volume 45, Issue 1, January 2009, Pages 20–34.

[13] A. K. Jain, M. N. Murty, and P. J. Flynn. 1999, Data clustering: a review, *ACM Comput. Surv. 31, 3 (September 1999), 264-323.*

[14] Lin, Chin-Yew and E.H. Hovy 2003, Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics, *In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, May 27 - June 1, 2003.*