ROBUST MACHINE LEARNING TECHNIQUES FOR DOCUMENT SUMMARIZATION

Feny Mehta[1], Sanjay Bhanderi[2], Arindam Chaudhuri[3]

[1]*PG Student, Computer Engineering, MEFGI, Gujarat, India*
[2]*Assistant Professor, Computer Engineering, MEFGI, Gujarat, India*
[3]*Associate Professor, Computer Engineering, MEFGI, Gujarat, India*

ABSTRACT

Currently huge amount of data is available on the internet which is increasing exponentially day by day. It becomes time consuming and tedious job to search a specific topic from the heap of information available. Document summarization is the key solution to the above stated problem. It refers to reducing the size of the document still preserving the main information of it. . Now to summarize the data using the computer program or algorithm is called the automatic document summarization. Abstractive and Extractive are the two main automatic document summarization techniques. To summarize the data mainly there are two steps, pre-process the data and process it, where in pre-processing the data is cleaned removing unwanted data and in processing various techniques are applied to summarize the data. Data summarization has various real life applications and is very useful for everyday life. This paper gives the hybrid approach for Multi-document Summarization where initially the documents are clustered using the effective grouping through advance similarity measure which also considers the dissimilarity of each document with every other document in the corpus, then the extraction technique is used for sentence extraction as they are reordered according to their weights obtained. Lastly the summary is effectively generated for each cluster.

Keyword: *Multi-document Summarization, clustering, extraction of sentences*

1. INTRODUCTION

Data mining is termed as computer-assisted process of digging through and analysing enormous sets of data, extracting the meaning of the analysed data sets. Data mining tools predict behaviours and future trends, allowing businesses to make proactive, knowledge-driven decisions, reduce time consumption for resolving some business query, search data base for hidden pattern, finding predictive information.

Text Mining is referred to as mining of textual data .It is the sub-domain of data mining. Text mining can be Text mining usually involves the process of structuring the input text deriving patterns within the structured data, and finally evaluation and interpretation of the output.

Text summarization can be of two types,

1) Abstractive method:

Abstractive methods builds an internal semantic representation and then uses Natural Language Processing (NLP) to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original.

2) Extractive.

In extractive summarization selection of a subset of existing words, phrases or sentences in the original text to form a summary is done.

Due to the exponential growth of textual information available on the web, it becomes time consuming for the users to get desired information from the huge text corpus. If the users are provided with the summary of text available without losing its important information, it would be of a great help. Text summarization is a process of reducing the size of a text while preserving its information content. Also Text summarization can be defined as representing a subset of data which contains the information of the entire set. Moreover automatic document summarization creates a summary or abstract of the entire document by selecting the most informative sentences from the document automatically through computer techniques or algorithms. Automatic text summarization is a process that takes a source text and presents the most important content in a condensed form in a manner sensitive to the user or task needs. Let us consider the example of news articles. Everyday millions of news articles are made available on the internet. Now there may be thousands of articles for a single event or incident occurred. If a common man wants to read all the information about a single event it will be a time consuming job to read each and every article to get information. Every article gives some similar and some unique or dissimilar information about a single event. If the most similar information can be extracted and represented for each event it would be easier for users. For this document summarization is very useful. Finding similarity and unique information either from a single document or multiple documents and extracting the sentences giving brief idea about the event.

2. RELATED WORK

Most of the approaches in the field of multi-document summarization use the supervised approach for the grouping of similar documents. But the main disadvantage here is that the classifier needs to be given the training dataset every time for different data set. Hence the unsupervised approach is to be used. But even though some researchers are using clustering techniques, they would use the traditional techniques for it. Here the main challenge is to find the robust and effective technique for clustering so that the documents can be clustered effectively. Now to summarize documents there is a need to find an approach so that the summary generated would be precise, accurate. There is a need to find a sentence extraction technique through which the sentence extracted would generate a summary more accurately as the ones generated by humans.

In this paper the following issues are addressed. The supervised techniques are used for grouping of the corpus which makes the system domain dependent .Even if the unsupervised techniques are used for the clustering of the documents; the traditional clustering algorithms are used which are of lower accuracy. To generate the summary the sentences are extracted on the basis of its relevance to the heading, position of the sentence, sentence length etc. Hence the system would be domain dependent.

In (2015) Tanmay Basu et al.[1] introduced a hybrid document clustering technique by combining a new hierarchical and the traditional k-means clustering techniques. A distance function is proposed to find the distance between the hierarchical clusters. Initially the algorithm constructs some clusters by the hierarchical clustering technique using the new distance function. Then k-means algorithm is performed by using the centroids of the hierarchical clusters to group the documents that are not included in the hierarchical clusters.

In (2015) Yogesh kumar Meena et al.[2] proposed a feature priority based filtering method for summarization where sentences are filtered using tf-isf scores, named entities and proper nouns. After POS scores to the sentences, the sentences are arranged in the decreasing order of their scores; the first sentence is taken as it is for the summary at first position.

In (2014) Sara Botelho Silveira et al.[3] described an approach that uses lexical and semantic, both sentence reduction techniques. Summarization is done in two phases, first clustering by similarity and then clustering by keywords. The sentences are ordered on the basis of tf-idf scores. Sentence reduction is performed by removing specific sentential constructions conveying less relevant information to the summary. The three main algorithms proposed for it are main clause, blind removal, and best removal. The final step of this algorithm determines if the new reduced sentence can replace the former sentence based on the sentence score, finally providing improved summary.

In (2013) Mohamed Abdel Fattah[5] proposes an approach that uses statistical tools to improve content selection in multi-document automatic text summarization. It uses the trainable summarizer, taking into account several features, which are then used in combination to construct text summarizer model. For final summary to be generated the hybrid model of maximum entropy, naive Bayesian classifier, support vector machine is proposed that ranks the sentences on the order of importance.

The rest of the paper is organized as follows: section 3 contains the proposed system, 4 experimental results, 5 Conclusion.

3. PROPOSED METHOD

The proposed framework has two major modules, clustering of the corpus, extraction of the sentence for summary generation.

The figure 1 shows the detailed view of the proposed framework along with their flow interactions.

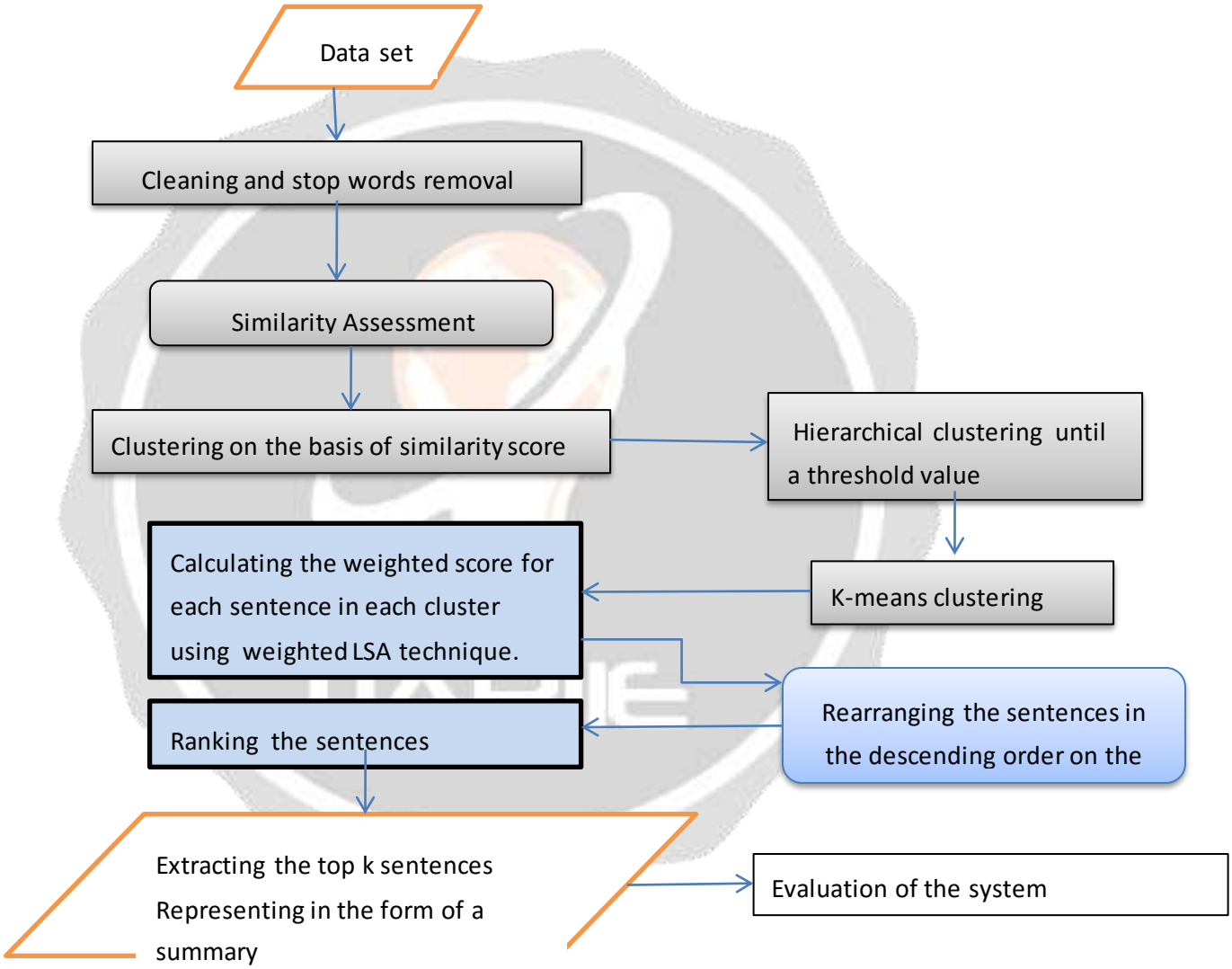

Fig 1: Proposed framework

Initially the dataset needs to be cleaned (i.e.) remove unwanted characters, hyperlinks etc. and to be given as input for the clustering module. To understand the working of the both the modules consider the below given figures.

3.1 Clustering Module

Initially each document is converted into vectors through vector space model. Its term frequency and inverse document frequency are calculated. After the TF-IDF weighting the cosine similarity scores of each sentence with every other sentence is calculated through the standard cosine similarity formula. After the scores are computed the Extensive similarity is to be generated through the following computation of cosine scores obtained.

The similarity between two documents is determined by checking their distances with every other document in the corpus, if they have a minimum content similarity.

Formula: The similarity measure is named as *Extensive Similarity*[1]. It is defined as,

$$ES(d_i, d_j) = N - \sum_{k=1}^{N} |\, dis(d_i, d_k) - dis(d_j, d_k) \,| \quad \text{if } dis(d_i, d_j) = 0$$
$$= -1 \quad \text{otherwise}$$

Where,

$$dis(d_i, d_j) = 1 \quad \text{if } \rho(d_i, d_j) \leq \theta$$
$$= 0 \quad \text{otherwise}$$

Here ρ is a similarity measure. In the context of text data it is assumed as cosine similarity measure.

Each document is assumed as a cluster initially. These singleton clusters are taken as an input for the hierarchical clustering. The hierarchical clustering merges these clusters on the basis of the extensive similarity score until predefined value alpha and hence the baseline clusters would be formed.

After this step there would be some singleton clusters left which are not the baseline clusters, for it the k-means clustering algorithm is to be used so that those single documents can be merged with baseline clusters with which they have the highest similarity score.

Thus effectively the documents would be clustered using this hybrid clustering algorithm which is a combination of hierarchical clustering technique and k-means clustering algorithm. Here the numbers of clusters are determined automatically.

3.2 Summary Generation Module

After the effective grouping of the documents the second step is to extract the important sentences from the cluster and give a precise summary.

There are two steps in this module. Initially the first step uses the modified Latent Semantic Analysis which uses the Singular Value Decomposition (SVD) for finding out the semantic meaning. This step unlike the existing LSA methods uses the Global weighting, local weighting scheme and the neighbour weight for the sentence selection process.

The formula for it is given as follows

$$a = L(tij) * G(tij) + N(tij)$$

Where $L(t_{ij})$ is the Local Weight for term in sentence, $G(t_{ij})$ is the Global Weight for term in the whole document, $N(t_{ij})$ is the Neighbour Weight of term in sentence. This method uses the term based selection method.

Now the above mentioned process produces the representative sentences for each clusters formed. Now every sentence obtains a score based on the weighting scheme mentioned above through SVD.

The sentence scores are then sorted into the descending order. The second step of summary generation module is to select the representative sentences according to the weighted score and then reorder them. The top k sentences are selected and represented in the form of the summary.

Due to the above two techniques, a hybrid approach obtained gives the readable and understandable summary sentences with higher precision.

Here a hybrid approach of weighted LSA and K-means summarization is taken, which provides a better summary.

The figure 2 gives the flow chart of the summary generation

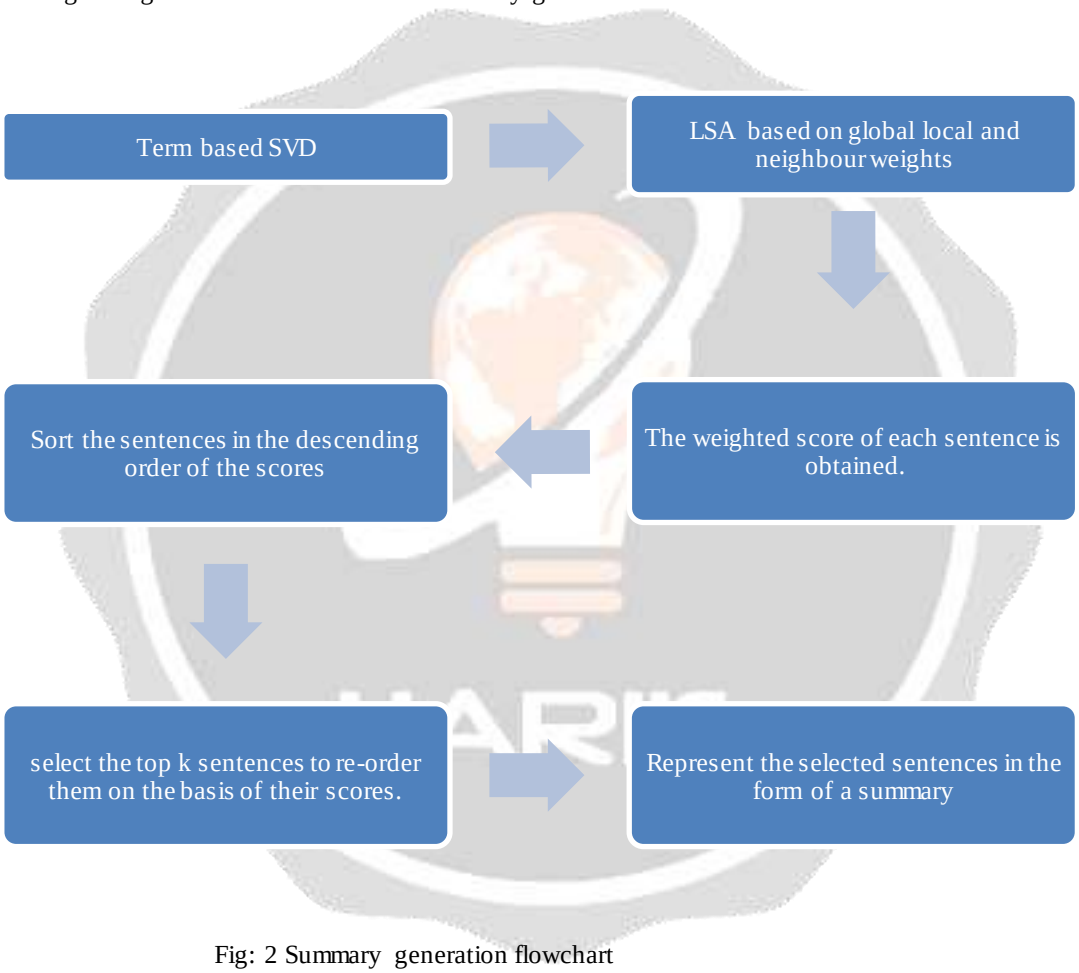

Fig: 2 Summary generation flowchart

For a single cluster formed all the sentences of all the documents in that cluster are considered.

Now the above mentioned module is run for all the clusters formed. The results are obtained cluster wise. That is the sentence are extracted cluster wise. So the summary formed gives the generic idea about the information contained in all the documents in a single cluster.

Hence the final output would be cluster wise summary for the documents contained in the database.

4. EXPERIMENTAL RESULTS

The dataset contains the news articles from Times of India news website. The dataset consists of 30 news articles document from the website. The documents are in the .txt format. The other data set contains news articles from the Google news website and from the Hindustan Times. The other data set contain the articles from blogs .The main aim is to subject the proposed system to different text dataset so that the robustness can be checked for the proposed system.

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (golden) summary or translation.

This toolkit is widely used by the researchers for the better evaluation of the system as it automatically computes the scores of the system.

Now the rouge2.0-distribution has been used here for the evaluation purpose. This toolkit gives the Average Precision, Average Recall and Average F-score for the system. The unigram and bigram are calculated and the ROUGE-1 and ROUGE-2 scores are obtained.

Now the evaluation of the proposed system is done using this toolkit. The results of the other existing systems are also taken and evaluated using toolkit for the better comparison of the proposed system.

Below fig shows the chart of the scores obtained for the evaluation of performance of the proposed system.

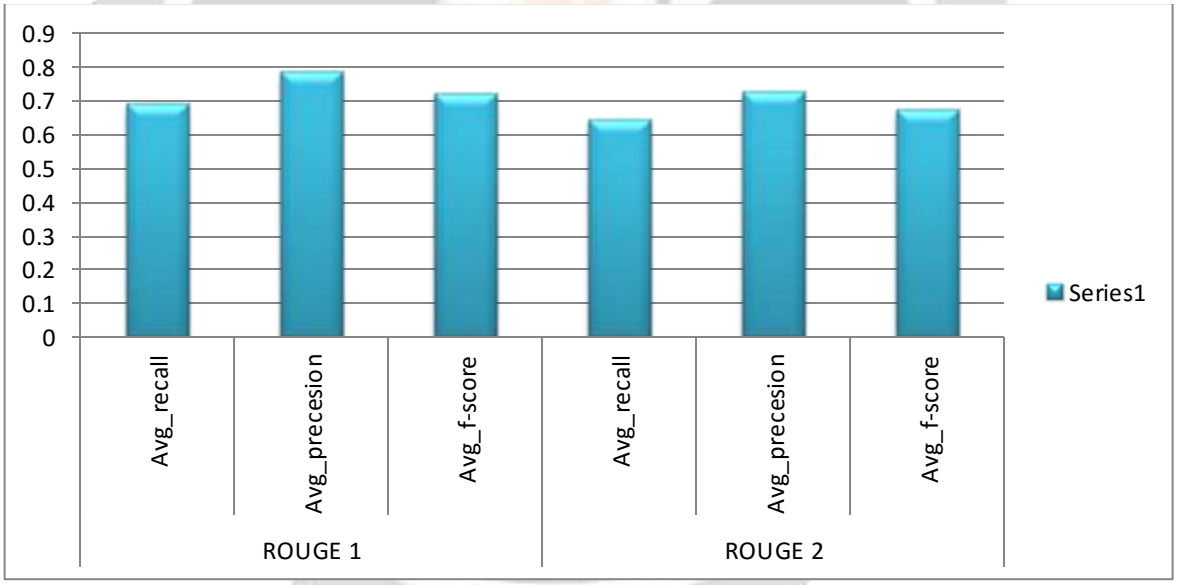

Fig 3 Scores for the Proposed System

The results of the proposed system are compared with the results of existing solutions. For the same purpose the existing algorithms LSA and LSA modified are run on the database taken. The results obtained from these algorithms are taken for the evaluation purpose. The summaries obtained from them are evaluated using the ROUGE toolkit.

The multiple reference summaries are taken with whom al the summaries generated by the considered systems are compared and the toolkit gives the average scores are obtained.

Consider the following tables which give the result of the ROUGE-1(unigram), ROUGE-2(bigrams) for the proposed system along with the scores obtained for the existing solution.

The table given below contains the ROUGE-1 scores. For comparison purpose the previous existing techniques are taken. Now over here the TASK are the different clusters formed for different databases taken for the evaluation of the proposed system. Hence here the result for 7 clusters of different database containing multiple documents is taken and defined as task. TASK1, 2, 3 contains the summary generated from the Times of India dataset, TASK 4, 5 contains the summary generated from the GOOGLE articles dataset, while TASK 6, 7, contains the BLOGS dataset.

Consider the following table for the ROUGE-1 scores of the existing solutions along with the scores of the proposed system

Table: 1 Comparison of Rouge-1 scores.

ROUGE-1									
	LSA			MODIFIED LSA			PROPOSED SYSTEM		
	Avg_R	Avg_P	Avg_F-Score	Avg_R	Avg_P	Avg_F-Score	Avg_R	Avg_P	Avg_F-Score
TASK1	0.5693	0.66162	0.61201	0.5558	0.58597	0.5705	0.88342	0.80196	0.84072
TASK2	0.5209	0.59326	0.55473	0.8044	0.56583	0.66436	0.62661	0.6943	0.65872
TASK3	0.1896	0.73077	0.30106	0.9402	0.64699	0.76651	0.51465	0.8308	0.66085
TASK4	0.6083	0.55732	0.58171	0.8424	0.51688	0.64065	0.67831	0.48522	0.56575
TASK5	0.3567	0.60784	0.44957	0.5415	0.72194	0.61884	0.79846	0.83708	0.81731
TASK6	0.456	0.6851	0.54754	0.5542	0.79864	0.65435	0.66438	0.89838	0.76745
TASK7	0.9034	0.75787	0.86226	0.4742	0.91489	0.62463	0.66638	0.87821	0.75777

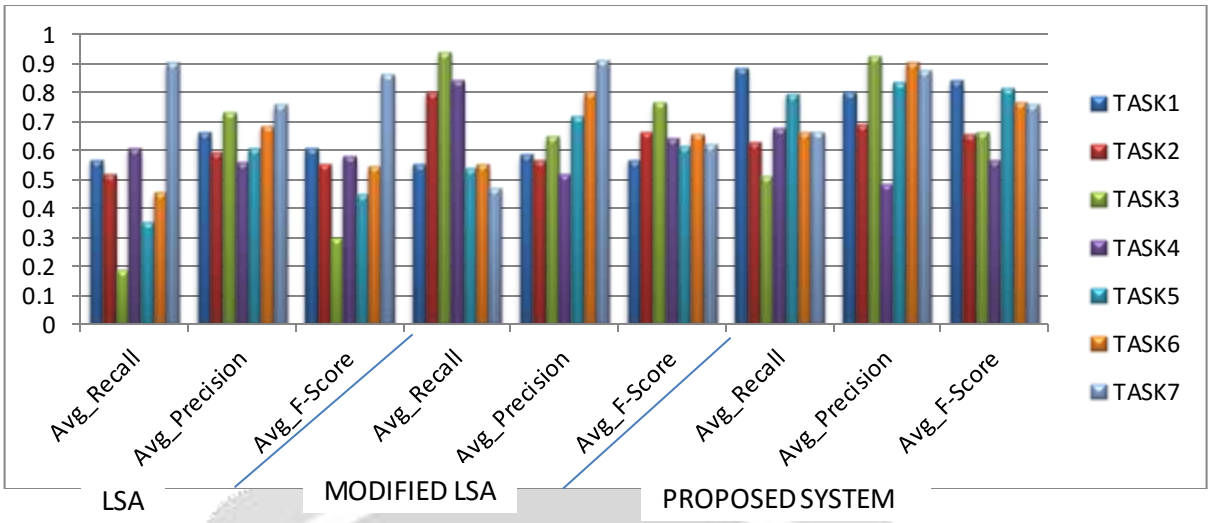

Fig: 4 Comparison of ROUGE-1 Scores

Similarly the ROUGE-2 scores for the TASKS are given below for the comparative analysis.

Table: 2 Comparison using ROUGE-2 scores

ROUGE-2									
	LSA			MODIFIED LSA			PROPOSED SYSTEM		
	Avg_R	Avg_P	Avg_F-Score	Avg_R	Avg_P	Avg_F-Score	Avg_R	Avg_P	Avg_F-Score
TASK1	0.4746	0.5602	0.51387	0.42576	0.4624	0.44334	0.87219	0.80738	0.83853
TASK2	0.3387	0.394	0.36424	0.70815	0.5	0.58614	0.54738	0.60541	0.57493
TASK3	0.108	0.5413	0.18004	0.89264	0.6287	0.73776	0.48442	0.90466	0.63097
TASK4	0.5016	0.4667	0.48349	0.76007	0.4693	0.5803	0.56018	0.41579	0.4773
TASK5	0.2269	0.3878	0.28626	0.47192	0.6309	0.53995	0.65358	0.89785	0.75649
TASK6	0.3214	0.5124	0.39504	0.45315	0.6912	0.54743	0.62758	0.84061	0.71864
TASK7	0.9858	0.7581	0.85712	0.43985	0.8736	0.58511	0.76171	0.81977	0.78967

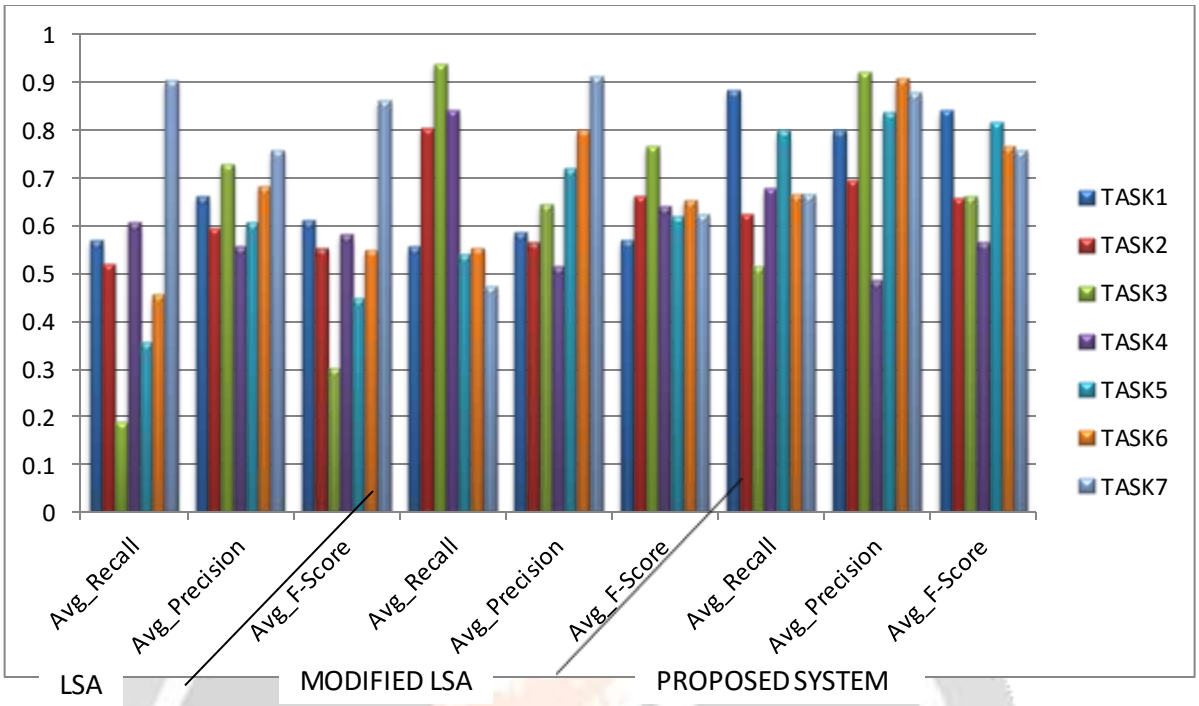

Fig: 5 Comparison using ROUGE-2 scores.

Now the average score comparison for the ROUGE-1 AND ROUGE-2 scores for all the 3 methods is shown in the below table.

Table 3: Comparison of Average ROUGE results

	ROUGE 1			ROUGE 2		
	Avg_Recall	Avg_Precesion	Avg F-score	Avg_Recall	Avg_Precesion	Avg F-score
LSA	0.528689	0.656254	0.558411	0.4224157	0.517221	0.440086
MODIFIED LSA	0.673247	0.708734	0.698549	0.593077	0.698026	0.62429
PROPOSED SYSTEM	**0.690316**	**0.789747**	**0.724081**	**0.6438629**	**0.7259243**	**0.67379**

The below shown figure is the bar-chart representation of the above scores for better understanding.

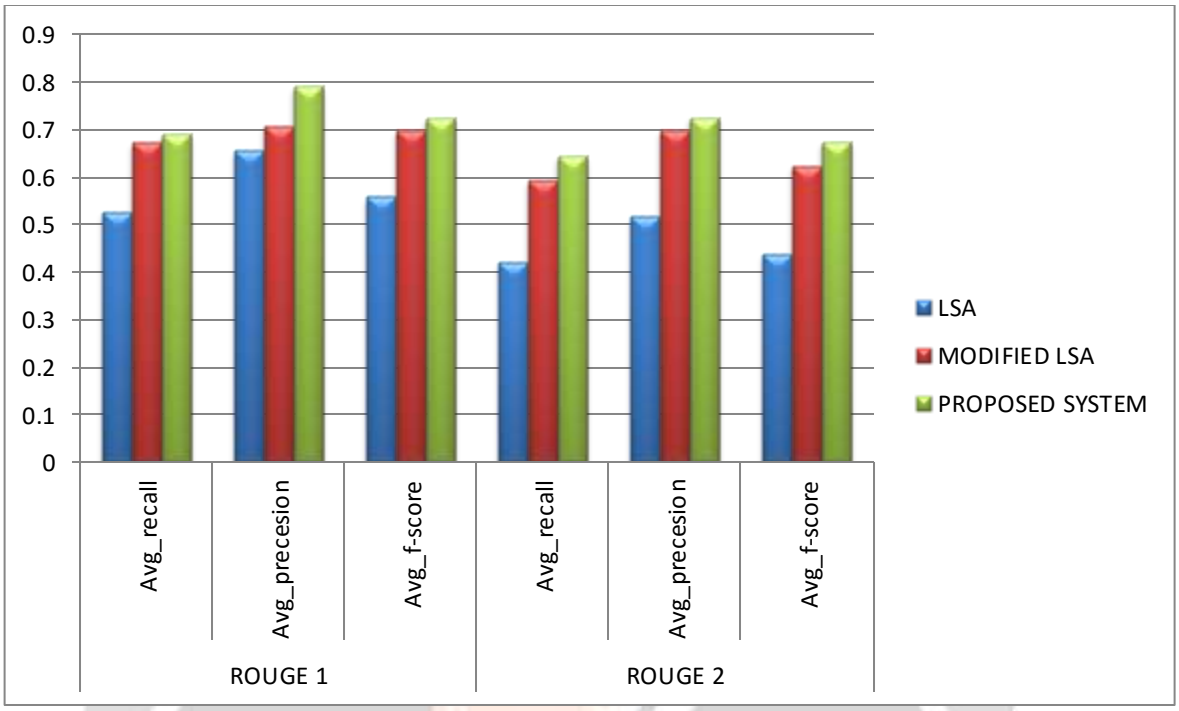

Fig: 6 Comparison of Average ROUGE scores

From the above comparison it can been seen that the proposed system outperforms the existing methods due to the reordering and re-ranking of the sentences on the basis of the weighted scores of the sentences obtained.

5. CONCLUSION

The Document summarization deals with providing the abstract or the main idea of the whole document. But the literature survey shows that most of the systems available are domain dependent and are not robust, moreover the sentence extraction is dependent on centroid value, position of sentence, its relevance with the title etc. Also the present multi-document summarizers need to be given the training data set to classify them and then summarize the documents. Hence the proposed uses a clustering algorithm, where the training data set is not required and it can cluster them according to the data set. The proposed summary generation algorithm increases the readability of the summary. Moreover the proposed system generates a summary which can be applicable on any text data and is a generic summary making the system domain independent.

The future work can be done on the dynamic selection of the number of sentences to produce the generic summary. Moreover the work can be done to apply the system on data other than text. The proposed work can be enhanced in the terms of abstraction and combining the NLP concepts for the generation of ideal human like summaries.

6. REFERENCES

[1] T. Basu and C. A. Murthy, "A similarity assessment technique for effective grouping of documents," Inf. Sci. (Ny)., vol. 311, pp. 149-162, 2015.

[2] Y. Ouyang, W. Li, R. Zhang, S. Li, and Q. Lu, "A progressive sentence selection strategy for document summarization," Inf. Process. Manag., vol. 49, no. 1, pp. 213-221, 2013.

[3] Y. K. Meena, D.Goplani, " Feature priority based sentence filtering method for extractive automatic text summarization," International Conference on Intelligent Computing, Communication and Convergence, Procedia Computer Science, Volume 48, pp. 728-734, 2015.

[4] Sara Botelho Silveria, Antonio Branco, "Sentence reduction algorithms to improve Multidocument summarization," 5th int. conf. on agents and artificial intelligence, pp. 261-276, 2014.

[5] Dey and D. Majumdar, "Fast Mining of Interesting Phrases from Subsets of Text Corpora," 17th int. conf. on Extending Database Technology, pp. 193-204, 2014.

[6] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han, "Mining Quality Phrases from Massive Text Corpora," Proc. 2015 ACM SIGMOD Int. Conf. Manag. Data, pp. 1729-1744, 2015.

[7] M. Fattah, "A hybrid machine learning model for multi-document summarization," Appl. Intell., vol. 40, no. 4, pp. 592-600, 2014.

[8] W. Chuang and J. Yang, "Extracting sentence segments for text summarization: a machine learning approach," Proc. 23rd Annu. Int. ACM , pp. 152-159, 2000.

[9] M. D. E. Buenaga, "Multidocument Summarization : An Added Value to Clustering in Interactive Retrieval," vol. 22, no. 2, pp. 215-241, 2004.

[10] M. Mendoza, S. Bonilla, C. Noguera, C. Cobos, and E. León, "Extractive single-document summarization based on genetic operators and guided local search," Expert Syst. Appl., vol. 41, no. 9, pp. 4158-4169, 2014.

[11] J. Chen and H. Zhuge, "Summarization of scientific documents by detecting common facts in citations," Futur. Gener. Comput. Syst., vol. 32, pp. 246-252, 2014.

[12] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," Inf. Process. Manag., vol. 40, no. 6, pp. 919-938, 2004.

[13] J. P. Mei and L. Chen, "SumCR: A new subtopic-based extractive approach for text summarization," Knowl. Inf. Syst., vol. 31, pp. 527-545, 2012.

[14] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," Expert Syst. Appl., vol. 36, no. 4, pp. 7764-7772, 2009.

[15] J. Atkinson, R. Munoz, "Rhetorics-based multi-document summarization," Expert Systems with Application, vol. 40, pp. 4346-4352, 2013.

[16] H. Hashimi, A. Hafez, and H. Mathkour, "Selection criteria for text mining approaches," Comput. Human Behav., vol. 51, pp. 729-733, 2015.

[17] C. Aggarwal and C. Zhai, Mining text data, vol. 4, no. 2(63). 2012.

[18] M. W. Berry and M. Castellanos, "Survey of Text Mining : Clustering , Classification , and Retrieval , Second Edition," 2007.

[19] R. Ferreira, L. de Souza Cabral, R. D. Lins, G. Pereira e Silva, F. Freitas, G. D. C. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, "Assessing sentence scoring techniques for extractive text summarization," Expert Syst. Appl., vol. 40, no. 14, pp. 5755-5764, 2013.

[20] T. Basu, C.A. Murthy Cues: a new hierarchical approach for document clustering, J. Pattern Recognit. Res. vol.8, no 1, pp.66–84,2013.

[21] Jiawei Han and Micheline Kamber, "Data Mining Concepts Techniques", 2011.

[22] Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram cooccurrence. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, May 27 - June 1.