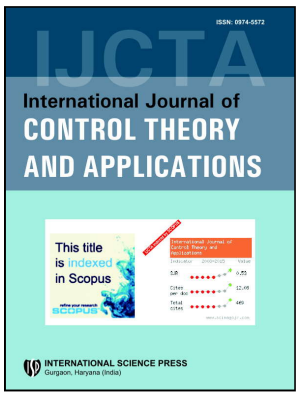

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 14 • 2017

Use of Hidden Markov Model to Enhance the Performance of Density Based Document Summarization

Supriya Anandrao Salunkhe[1] and Mrunal Bewoor[1]

[1] Department of Computer Engineering Bharati Vidyapeeth Deemed University College of Engineering, pune, Maharashtra, India, Emails: supriyasalunkhe43@gmail.com, msbewoor@bvucoep.edu.in

Abstract: In today's world use of online information increased as it is freely available. The retrieval of this information from available data, leads to a wide research in the area of automatic text summarization. It is necessary to apply the correct method to summarize the all available information. This solution is proposed for the Natural Language Processing (NLP) community.

Document summarization helps in mining data and delivering accurate data in time to the users. The system tries to attempt issues related to the data mining using various summarization methods.

Keywords: Hidden Markov Model, Density Based Algorithm, Document Summarization, Machine Learning.

1. INTRODUCTION

Now days, information is very important in whole world. More unstructured data involves during the internet searching, that is not possible for any user to read content of the any retrieved documents. So, it would be very necessary to discover method which will permit the user for extracting the correct approach by using retrieved documents. Summarizing all retrieved documents is useful to achieve the proposed objective. For this, new approach is introduced which will highlight necessary contents of the document. Summarization of documents are procedure of data mining which creates topic which will targets on minimizing the document size, while keeping in mind the important characteristics of our original document. Information overhead is critical problem in document summarization. This problem is generated by sharing the same topic by the multiple documents. So, document summarization is technique to overcome the above problem. There are different problems related to the document summarization which can be handled like matching of simple word as well as frequencies of word. This method doesn't search for correct similarity among words/or sentences within summary. Because it may happen that documents hold number of words which will explain similar event. It will considered relationship among words or sentences. Document summarization has two main steps. First is extracting information and another is ordering the sentence. This proposed system is focused on ranking as well as ordering. The way of arranging sentences combining which are extracted is known as ordering the sentences. Due to which it increases the summary readability. More research work is done for proposing the method for sentence ranking; these

documents may not consider the semantic aspects for word matching. Both statements has related to each other because the meaning of the both statements are same. Some types of the similarities like Syntactic, lexical and cosine similarity have been used for determining the relation among these statements.

2. LITERATURE SURVEY

Ms. Pallavi.D.Patil et al. *[1] Provides process of summarizing documents assumes a basic part in numerous applications. Text Summarization has been deliberate on keep hold of the essential data without concerning the archive quality.*

*AbimbolaSoriyan[2]*Summarization of documents help us to save information processing time as well as information meaning which is not predictable by user for using. Document summarization is supported by sentence compression by decreasing the summary candidate's length by maintaining candidate's applicable content, thus permitting inclusion space for further objects. Results of the paper does shows that adverbs, conjunctions of word, verb and adjectives and more are get deleted without losing sentence meaning.

P. Sukumar et.al. [3] Through fast growth of unstructured information by naturally may leads big problem to algorithm of text mining. This will recover information which is meaningful throughproficient way. But more amounts of data are easilyobtainable; sometime it is not possible to get essential or appropriate information on right time. Hence, for generating the topic summary we needs document summarization. Document verification mainly targets on the ranking and ordering of the sentences. Chronology, succulence, topical and precedence are some of the measures of the sentence ordering which the proposed system deals with.

3. EXPERIMENT AND RESULT

3.1. System Design

To provide linguistic knowledge to the computer system is the big challenge for natural language processing. The Figure 1 shows the overall document summarization.

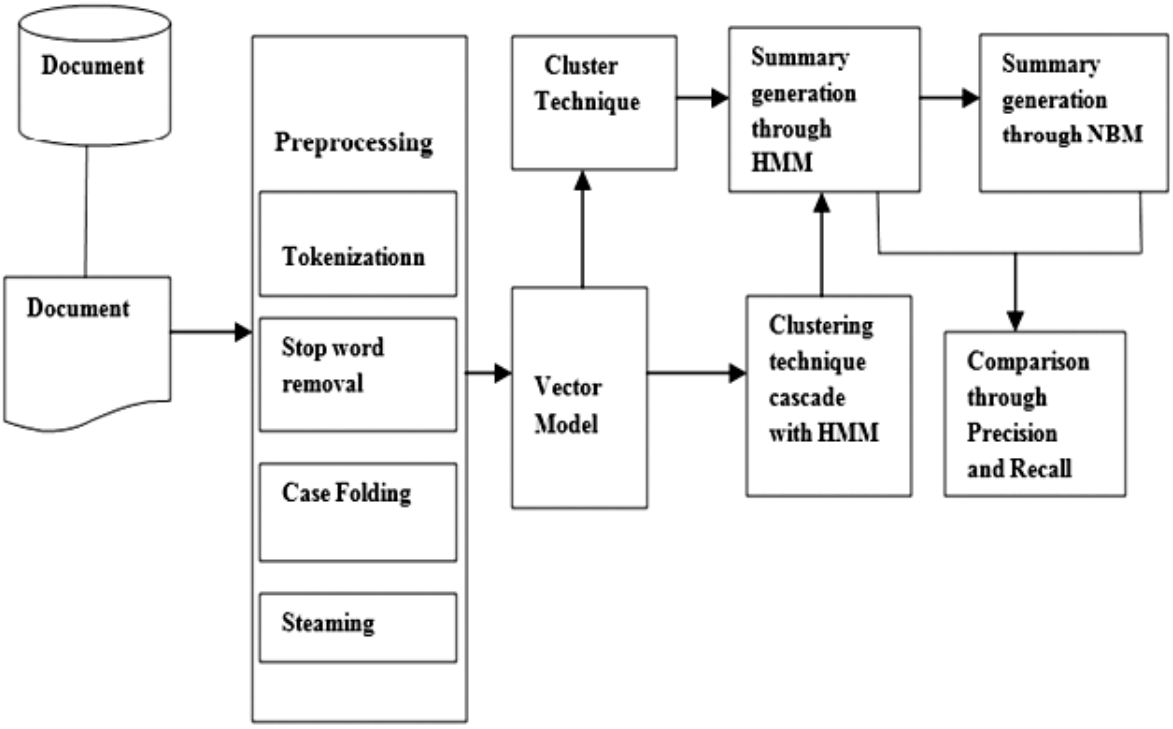

Figure 1: Proposed System Architecture

On the basis of such considerations, the algorithm uses a different color image multiplied by the weighting coefficients of different ways to solve the visual distortion, and by embedding the watermark, wavelet coefficients of many ways, enhance the robustness of the watermark.

3.2. Preprocessing

The tokens are generated from the preprocessing unit by giving the number of documents to this preprocessing unit. These documents are broken into smaller atomic parts these called as token. The word which doesn't explain the sentence meaning is called as stop word. This stop words are deleted to avoid the unwanted word processing.

3.3. Similarity Matrix Calculation

Calculating the weight of word for understanding the summary of the document is very significant. TF–ISF (Frequency–Inverse Sentence Frequency) is utilized for this calculation. The number of occurrences words in the document is known as Frequency. The percentage of containing words in the sentence is calculated by using the Inverse sentence frequency. Term Frequency is calculated as follows:

$$Term\ Frequency = \frac{nj}{\sum_k nk} \tag{1}$$

In above Equation,

nj is the number of occurrence of word j in summary

nk is the total words in summary.

Inverse Sentence Frequency is calculated as follows:

$$ISF = \log \frac{N}{ni} \tag{2}$$

In above Equation,

N represents total sentences in summary.

ni represents no. of sentences holding specific term

Term Weight is determined as follows:

$$Term\ Weight = TF * ISF \tag{3}$$

3.4. Mathematical Model

$T_1[D] = \sum_{ti \in n}[D]$

$T_2[D] = \sum_{ti \in n}[D]$

……

$T_n[D] = \sum_{ti \in n}[D]$

Where,

T1….Tn = Number of tasks

D = d1..... dn (No. of Documents)

Algorithm for Document Clustering

Input: Ranking score p of each sentence
Output: Clustered Document

1. Draw a graph which nodes is the point to be clusterd
2. Draw an edge from c to each p (point) do for each core point c around c ε-neighborhood
3. Put N on the graph node;
4. Delete N if it doesn't holds any core point
5. Choose a 'c' (core point) from N
6. Let X is node set which may reach from c by moving ahead.
   1. Make cluster contains ("X∪{c}")
   2. "N=N/(X∪{c})"
7. Repeat/Continue step

4. SUMMARIZATION MODEL

4.1. Hidden Markov Model

This model describes an approach that given a set of features computes an a-posterior probability that each sentence is a summary sentence. In contrast to a Naive Bayesian Model(NBM) approach, the HMM has fewer assumptions of independence. In particular, it does not assume that the probability that sentence i is in the summary is independent of whether sentence i1 is in the summary. Furthermore, the joint distribution for the features set, is used is the assumption used by naive Bayesian method. The three features are considered in the development of a Hidden Markov model for text summarization.

1. Position of the sentence in the document. This feature is built into the state-structure of the HMM.
2. Number of terms in the sentence. The value of this feature is

   o1(i) = log(number of terms + 1):
3. How likely sentence terms are, given the document terms

   o2(i) = log(P r(terms in sentence i | D)) .

The probability that the next sentence is included in the summary will differ, depending on whether the current sentence is a summary sentence or not. The Hidden Markov Model allows such differences with marginal additional cost over a simple Bayesian classifier.

5. RESEARCH EVALUATION

5.1. Experimental Setup

The experimental setup used for this system is given as below. Windows 7 OS is used on Intel Pentium Dual Core Processor and 2 GB main memory. The hard disk used for this is of 320 GB. This system is implemented using Java framework.

5.2. Analysis of work

The quality of summary obtained is measured using following parameters :

1. Precision –

   Precision is the fraction of the documents retrieved that are relevant to the user's information need.

$$precision = \frac{\left|\{relevant\ documents\} \cap \{retrieved\ documents\}\right|}{\left|\{retrieved\ documents\}\right|}$$

   That is, the fraction of the retrieved documents which is relevant.

2. Recall –

   Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$recall = \frac{\left|\{relevant\ documents\} \cap \{retrieved\ documents\}\right|}{\left|\{relevant\ documents\}\right|}$$

   That is the fraction of the relevant documents which has been retrieved.

The paragraph is given as input to the system which was further processed by applying summarization technique. The objective of this work is to produce the accurate summary. The result obtained by the system is the summary of that input paragraph. The analysis of result shows that some sentences of the input paragraph are not summarized and some sentences are added into summary those can be omitted. The system is tested for 25 various inputs and following graph indicate analysis of these inputs.

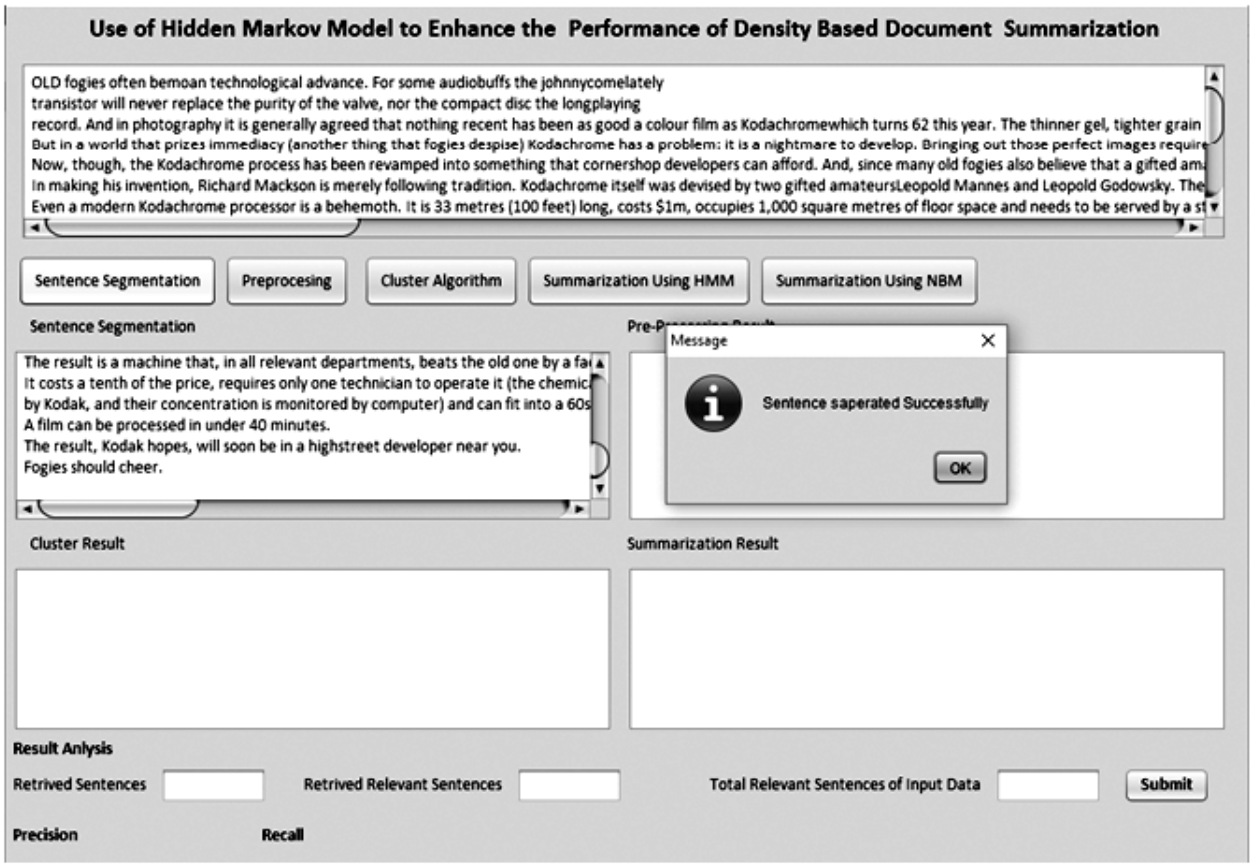

Figure 2: Document Segmentation

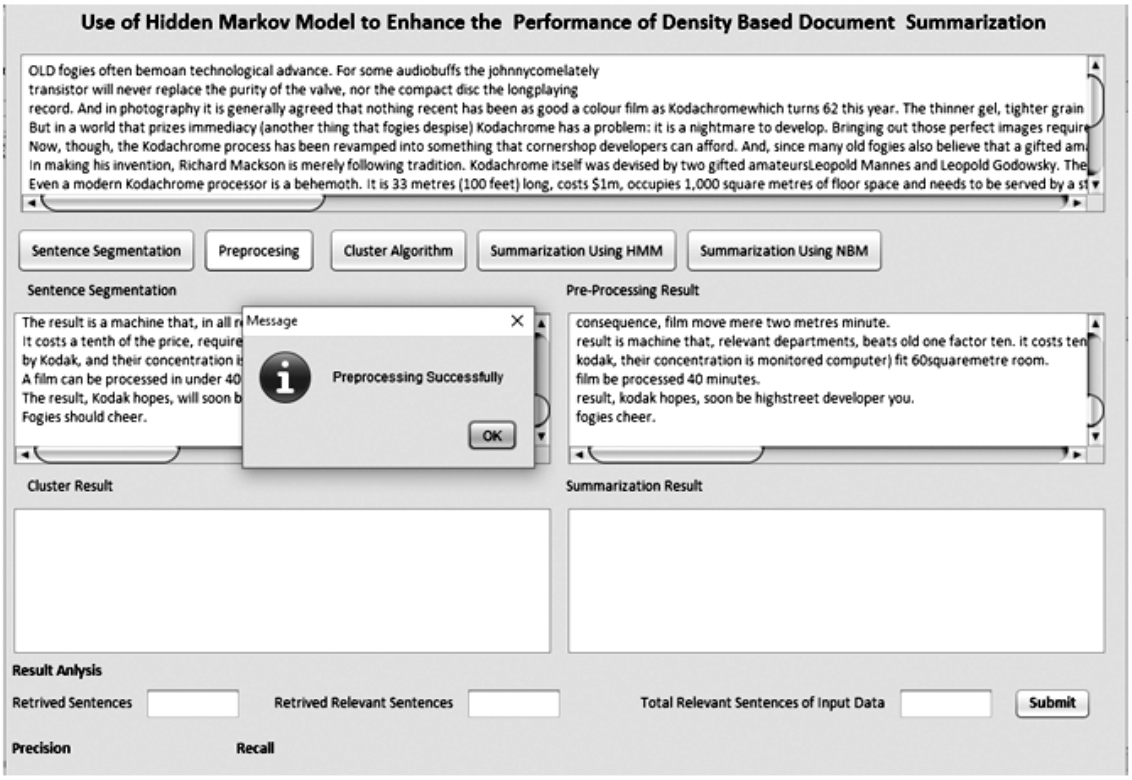

Figure 3: Document preprocessing

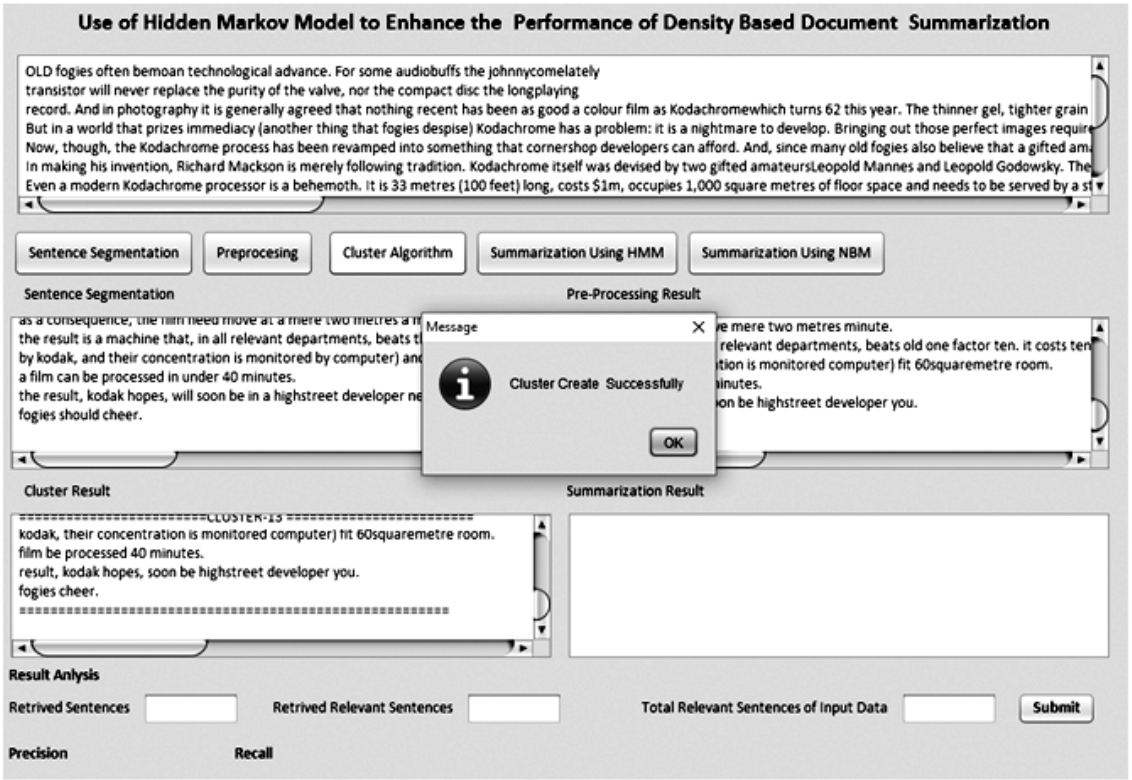

Figure 4: Document Clustering

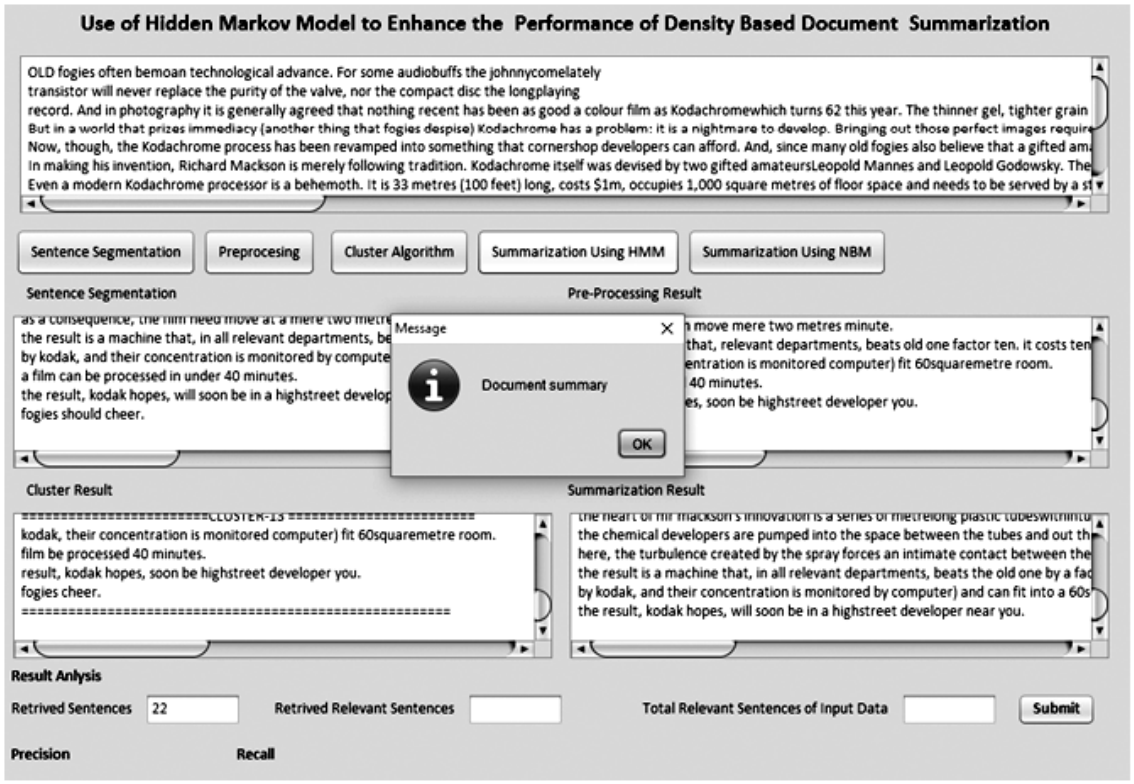

Figure 5: Document Summary Using HMM

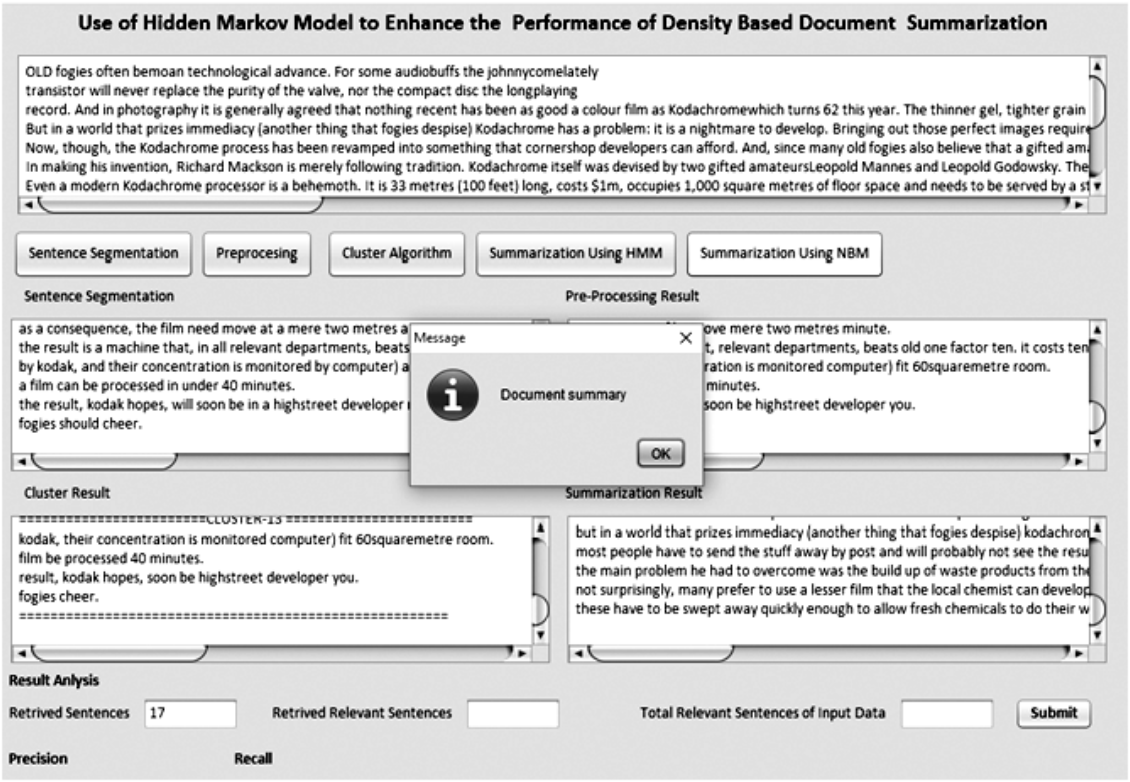

Figure 6: Document Summary Using NBM

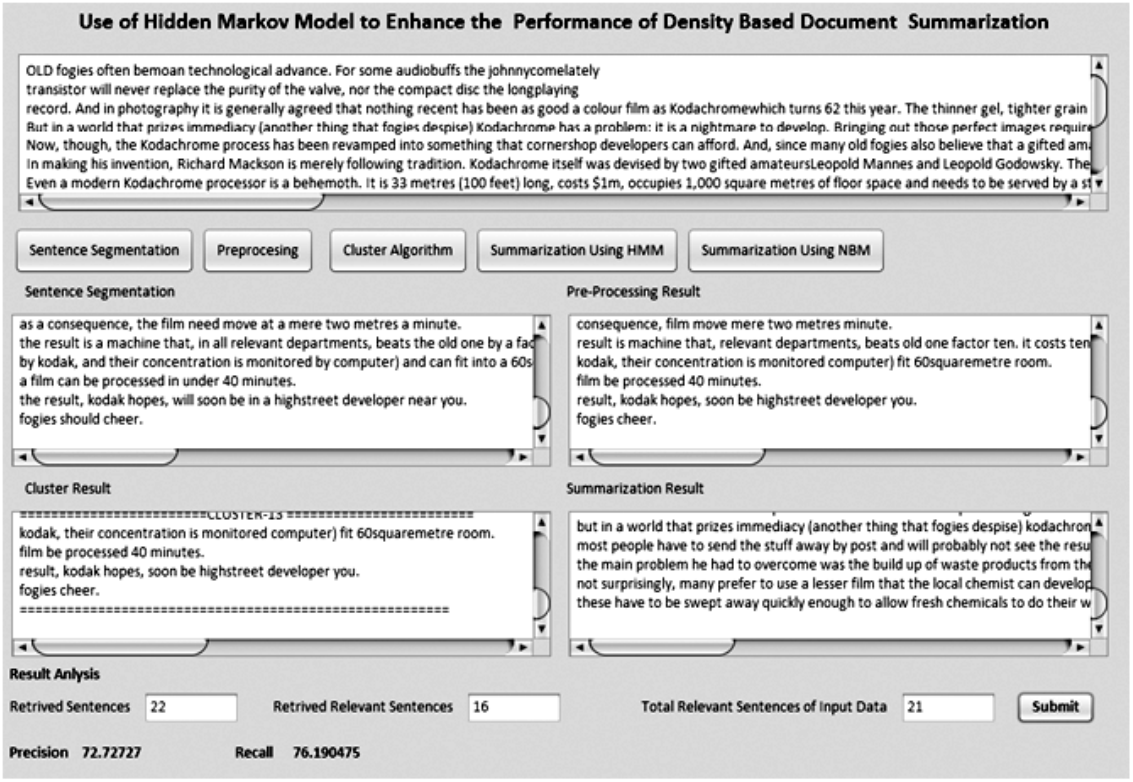

Figure 7: Analysis of summary

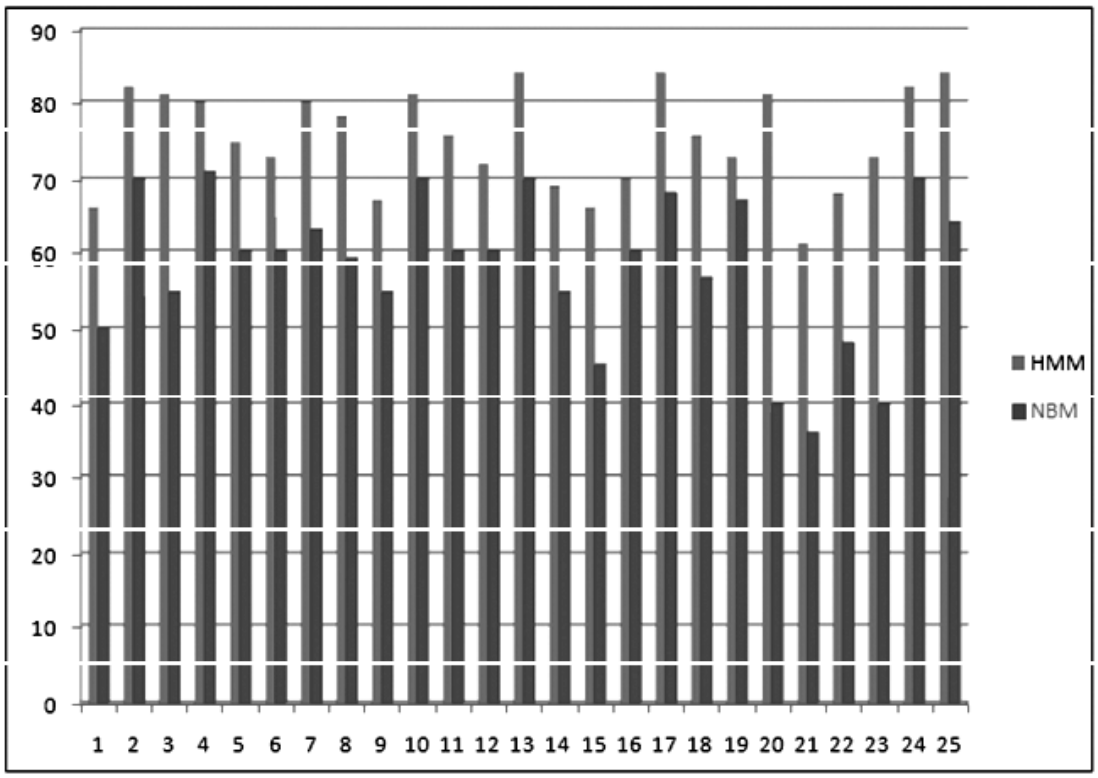

Figure 8:HMMvs NBM of Twenty five input Documents.

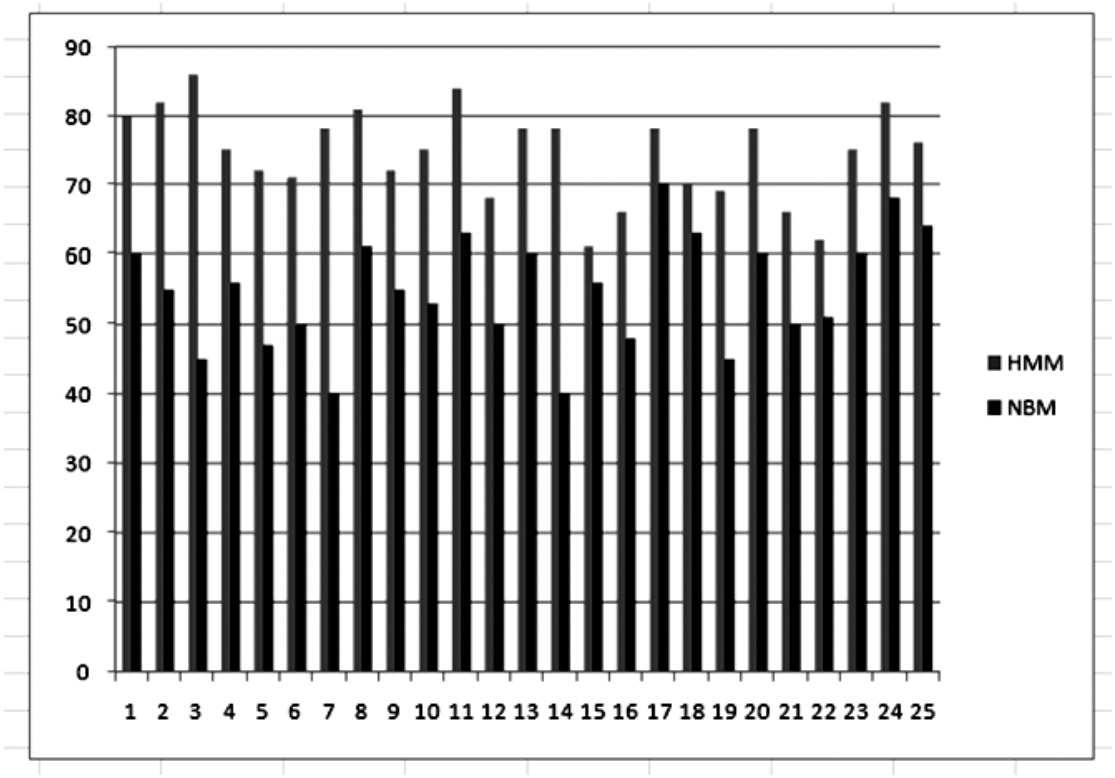

Figure 9: Recall of HMM vs NBM Twenty five input Documents

4. CONCLUSION

This paper presents how Hidden Markov Model can be used to enhance the performance of density based document summarization. In the various dimensions the document summarization methods are improved. These implemented method is applied on the document from various domains. The focus of research work is to reduce the information overload at the time of information retrieval. It is especially in the field where large amount of information like World Wide Web, hospital information and biomedical articles is used. Compared with different document summarization technique available and their evaluations compared with human summarization are still not satisfactory. Most document summarization is extractive and while human summarization is mostly abstractive that needs NLP to construct sentences.. This is one of the major issues in abstractive document summarization. This issue will be tackled in the future work. This research work could be integrated in systems like [7,9] and have scope of enhancement

REFERENCES

[1] Ms.Pallavi.D. Patil, P.M. Mane M.E Scholar, Assistant professor Department of Computer Science and Engineering Dnyanganga College of Engg. & Research,Narhe, Pune, India," Improving the Performance for Single and Multi-document Text Summarization via LSA & FL." International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 4, Jul-Aug 2015.

[2] Abimbola Soriyan ,Theresa Omodunbi , Dept. of Comp. Sci. & Engineering Obafemi Awolowo University, Nigeria, "Trends in Multi-document Summarization System Methods,"International Journal of Computer Applications (0975– 8887) Volume 97– No.16, July 2014

[3] P.Sukumar, K.S.Gayathri ," Semantic based Sentence Ordering Approach for Multi-Document Summarization" International Journal of Recent Technology and Engineering (IJRTE)ISSN: 2277-3878, Volume-3, Issue-2, May 201471 Published By:Blue Eyes Intelligence Engineering& Sciences Publication Pvt. Ltd.

[4] Aurelien Bossard, Christophe Rodrigues ," Combining a Multi-Document Update Summarization System", in Laboratoire d'informatique de Paris Nord, CNRS UMR 7030 Universite Paris 13, 93430 Villetaneuse, FRANCE e-mail: firstname.lastname@lipn.univ-paris13.fr.

[5] M. S. Bewoor, S. H. Patil, "the Performance of Cluster-Based Document Summarization,"Associate Professor Computer Department, Bharati Vidyapeeth Deemed University College of Engineering Pune India

[6] Mark Stamp,"ARevealing Introduction to Hidden Markov Models,"Department of Computer Science, San Jose State University in August 31, 2015

[7] Kadam, Aniket D., et al. "Question Answering Search engine short review and road-map to future QA Search Engine." *Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015 International Conference on* IEEE, 2015.

[8] Kadam, Aniket D., et al. "Hybrid intelligent trail to search engine answering machine: Squat appraisal on pedestal technology (hybrid search machine)." *Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015 International Conference on*. IEEE, 2015.

[9] Todkar, Omkar, S. Z. Gawali, and Aniket D. Kadam. "Recommendation engine feedback session strategy for mapping user search goals (FFS: Recommendation system)." *Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on*. IEEE, 2016.

[10] Shinde, Snehal, et al. "A decision support engine: Heuristic review analysis on information extraction system and mining comparable objects from comparable concepts (Decision support engine)." *Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on.* IEEE, 2016.

[11] Salunkhe, Pramod, et al. "Hybrid machine translation for English to Marathi: A research evaluation in Machine Translation:(Hybrid translator)." *Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on*. IEEE, 2016.

[12] Joshi, S. D. "QAS." International Journal of Application or Innovation in Engineering & Management 3.5 (2014): 429-436.

[13] Kumar, Naveen, et al. "A Scalable Record Retrieval Methodology Using Relational Keyword Search System." Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. ACM, 2016.

[14] Salunkhe, Pramod, et al. "Hybrid machine translation for English to Marathi: A research evaluation in Machine Translation:(Hybrid translator)." *Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on*. IEEE, 2016.