TWO-TIER ARCHITECTURE FOR DOMAIN SPECIFIC DOCUMENT SUMMARIZATION USING PROBABILISTIC LATENT SEMANTIC ANALYSIS

Madhuri Singh[1], Farhat Ullah Khan[2]
[1]Department of Computer Science & Engineering, Amity University, Noida, India
[2]Department of Computer Science & Engineering, Amity University, Noida, India

ABSTRACT

In this research work we have proposed two-tier architecture for document summarization. This architecture minimizes the redundancy and boosts the information relevancy in the summary by applying Probabilistic Latent Semantic Analysis (PLSA) at two levels. It also enhances the summarizer's speed by using Incremental Expectation Maximization algorithm for PLSA learning rather than Expectation Maximization. It starts with collecting number of topical information from multiple news portals and applies PLSA for single document Summarization. At next level PLSA is applied again in order to produce final summary but this time for multiple-document summarization. Here the two-tier stands for single and multiple-document summarization. In this paper we have performed a summarization experiment and we have also given a brief report of our experimental results. We experimented with variety of documents for several times and observed that results are generated in a matter of few seconds and moreover, the quality of the solution is fairly improved. To validate our results we have used the ROUGE metrics. The validation results devise the effectiveness of the proposed architecture.

KEYWORDS: *Information Retrieval, Semantic space, Incremental EM, Extractive summary, PLSA.*

I. INTRODUCTION

Evolution of Internet has revolutionized the era of Information Technology. It has enabled numerous users to access information and other services instantly, irrespective of their location but superfluous in amount. If a person fires a query using search engine he gets a large list of pages even if he needs a two line solution. The searching of useful information from these pages consumes lot of precious time and downgrades the search engine's performance as well. Summarization of documents might be a solution to tackle this problem effectively. Document summarization is a technique of compressing the content size and still preserving the whole essence of the source content. [6] Summarization methods are widely divided into two categories – Abstractive Summarization and Extractive Summarization. Abstractive summarization needs the proper understanding of the input document similar to human brain to generate summary and is a very complex approach whereas Extractive Summarization does not need document understanding. It just picks the essential statements and paragraphs from the input document and fuses them together to restate the original content in fewer words.

Although numerous summarization techniques exist but the problem with those summarization approaches is that it focuses on minimizing the information redundancy from the summary. In process the relevant information gets ignored. The aim of these summarizers should be to extract relevant information and remove redundancies. To overcome this problem we have presented a two-tier architecture, which handles the process of extracting relevant information in effective manner.

The methods for summarization evolved in early fifties. The continuous research in this area has given birth to various popular techniques such as [1] Graph based techniques which involves

PageRank and Hyperlinked Induced Topic Search(HITS),Maximal Marginal Relevance (MMR) [1], LSA [1] and PLSA[1][2]. Although graph based techniques are popular but they suffer with a major drawback, that, the output gist contains only single topic i.e. the largest eigenvector is picked up as central topic which sometimes ignore the inclusion of crucial sentences in the summary and increases the redundancy. LSA overcomes this drawback by converting every sentence of a document in a Latent Semantic space before summary generation. But lack of algebraic foundation of LSA has made it less satisfactory. [4] Probabilistic Latent Semantic Analysis (PLSA) is a probabilistic version of the LSA and its statistical base is very strong because it is based on the Likelihood principle. The PLSA is more reliable than LSA and also covers various topics of the source document in the resultant summary.

In this paper we have used PLSA and discussed how our architecture improves the previous methods in following ways: In our first move, we apply PLSA in the context of single and multiple-document summarization both, taking in to account data redundancy and relevancy. Second, we represent only source documents in the semantic space which simply argues that, there is no need to represent the query into semantic space which reduces complexity involved in it. Third, we investigate how our model affects the summary quality by validating the summaries using ROUGE-L metric. The main motive of our proposal is to produce quick and high quality summaries.

The rest of the paper is organized as follows. Section II briefly covers the previous work. A brief PLSA description is given in section III. Section IV explains our proposed work. Sections V covers experimental results showing that our approach leads to improvement over basic PLSA and section VI contains conclusions & future work.

II. RELATED WORK

Handling Information overflow is a challenging task. Summarization is a way that makes information management easy. Summarization techniques started to evolve long time back. The popular summarization techniques used in fifties were Title, Cue, Location and Standard keyword methods [6]. The continuous development in this field brought techniques using HMM and clustering.

Probabilistic Latent Semantic Analysis also known as Probabilistic Latent Semantic Indexing was introduced by Thomas Hofmann in 1999 [4]. In 2001 T. Hofmann [5] published a research work showing how PLSA can be used for unsupervised learning and how tempered Expectation Maximization affects this process. He also showed various advantages of PLSA over LSA. Z. Sun et al. [9] described an event driven document selection method for information extraction that considers only domain specific documents and converts the various events in a document in the form of entities and relationships. Then various pattern based selection strategies are used to maximize information gain. F.L. Wang et al. [8] presented a hierarchical structure to arrange the various news stories and then Fractal summarization model is used to generate summary.

In 2008 a PLSA summarizer for single document extractive summarization was developed [1]. It uses standard Expectation Maximization for parameter estimation which speed downs the summarizer. In 2009 L. Hennig [2] proposed a query focused multi-document summarizer using PLSA which maps the query and the documents both in a latent semantic space. It trains the PLSA with historical summarization data and shows that PLSA is more suitable for capturing sparse information in a sentence than the LSI.

J. Xu et al. [3] used PLSA for human action recognition in videos in year 2009. They proposed an incremental version of EM and tested it on challenging human action datasets and found that incremental EM gives better action recognition and needs less iterations to execute in comparison to batch EM algorithm. K. Zhou et al. [10] proposed a slight modification of PLSA for learning with positive and unlabeled data. They have combined the unsupervised PLSA with some supervised information from user. It is very useful for required topic related document finding, even in the cases where very less positive examples are available. In single document summarization it is easy to find the sentence ordering from the input document but in case of multiple-documents it becomes very complex. An approach focused on the extraction of ordering of the sentences in case of multi-document summarization by combining the constraints from order of events and topic relatedness [12]. F. Zhuang et al. [8] presented a generative model using PLSA for multi-view learning. It models the co-occurrences of features and documents from different perspectives and follows the scheme of

co-training. A modified version of PLSA can improve the multi-document summarizer performance by integrating the clustering and summarization process together [13].

Along with the development of summarization techniques various summary evaluation methods came in to existence. One such popular method is summary evaluation using ROUGE. In 2004 C.Y. Lin [11] presented a paper that gave brief description of various ROUGE features. Research in this field is on but ROUGE has been considered as a successful evaluation method.

III. PLSA WITH INCREMENTAL EM

PLSA is probabilistic approach that assumes that a document covers number of topics or latent variables and each word in the sentence belongs to a latent variable with certain probability. It assumes following-

D is set of documents where $d \in D = \{d_1,d_2,\ldots,d_j\}$

W is set of words where $w \in W = \{w_1,w_2,\ldots,w_i\}$

Z is set of unobserved class variables where $z \in Z = \{z_1,z_2,\ldots,z_k\}$

Every document has a probability $p(d)$ associated with it and it belongs to a certain latent class z with probability $p(z|d)$ and every latent class generates a word with probability $p(w|z)$. The probability of every observation pair (d,w) can be defined as

$$p(d,w) = p(d)p(w|d) \text{ where} \quad (1)$$

$$p(w|d) = \sum_{z \epsilon Z} p(w|z)p(z|d) \quad (2)$$

Where z is given and word w and document d are conditionally independent. Using the frequency of word w in document d that is $n(d,w)$ the mixing components and mixing proportions can be determined by following formula-

$$l = \sum_{d \epsilon D} \sum_{w \epsilon W} n(d,w) \log P(d|,w) \quad (3)$$

For maximization of equation (1) in the presence of latent variables Expectation Maximization (EM) algorithm is used [4]. For single document summarization d represents a sentence s, z represents a topic and w represents a word [1]. [5] The two steps of EM algorithm are given below -

E-Step- It is used to calculate value of Posterior variable $p(z_k|d_i, w_j)$ using following formula-

$$p(z_k|d_i, w_j) \ \alpha \ p(w_j|z_k) \ P(z_k|s_i) \quad (4)$$

M-Step- It is used to update the value of posterior probabilities using following formulas-

$$P(z_k|d_i) \ \alpha \ \sum_{J} n(d_i, w_j) \ p(z_k|d_i, w_j) \quad (5)$$

$$P(w_j|z_k) \ \alpha \ \sum_{i} n(d_i, w_j) \ p(z_k|d_i, w_j) \quad (6)$$

$$P(z_k) \ \alpha \ \sum_{i,j} n(d_i, w_j) \ p(z_k|d_i, w_j) \quad (7)$$

Thus EM algorithm trains the PLSA. It starts with E-step and then goes to M-step. This alteration between two steps continues until the convergence is achieved. Since $p(z_k|d_i, w_j)$ in E-step, is updated using whole $P(z_k|d_i)$ and $P(w_j z_k)$ the PLSA training consumes lot of time. Hence we have employed Incremental EM in our work. [3] For human action recognition Incremental EM has been proved to be very effective. The main characteristic of Incremental EM is that it uses only a subset of previous data for updating various PLSA parameters.

IV. PROPOSED WORK

In our work we have conducted summarization at two levels. We have assumed that real time web crawler collects domain specific data. The pictorial view of our proposed architecture is given in Figure 1. We have performed summarization on the dataset collected from various news portals.

Before application of PLSA, each document is preprocessed which involves sentence splitting, stop word removal and Stemming. Once documents are preprocessed, word-sentence matrix is created for

each document and PLSA learning begins. Initially training takes places with random values and estimation parameters are calculated based on this initial data. But as the Incremental EM iterates to reach convergence the values of posterior probabilities becomes more accurate and is used to updates estimation parameters at next iteration. Once Incremental EM reaches its convergence, PLSA training ends and estimation parameters become more mature. By using these parameter values a score is generated for each sentence which describes the importance of the sentence in the document. These scores are generated [1] using following formula-

$$sentence\ score = \sum_{k} p(d_{i,}|z_{k,})\, p(z_k) \quad (8)$$

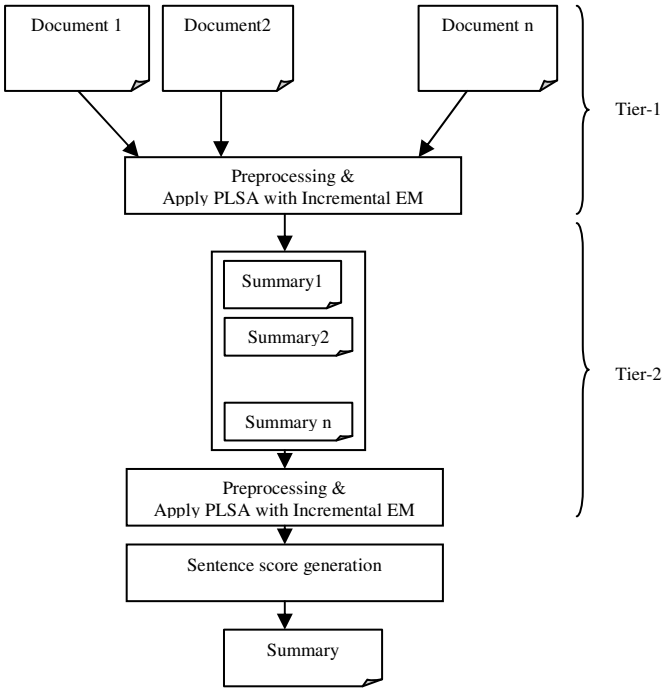

Figure 1. Proposed summarization architecture

After sentence score generation a threshold is selected empirically. Here first we calculate the average sentence score *avg*. The threshold *th* is calculated using following formula-

$$th = avg + avg/nos \quad (9)$$

Where *nos* represents the total number of sentence in the document. All sentences having greater score than the threshold *th* are kept in the summary and rest are discarded. As a result number of summaries is obtained. If there are *n* documents then *n* summaries are obtained. These summaries are obtained at tier-1 as a result of single document summarization. At tier-2 these *n* summaries are concatenated to form a single document. Now PLSA with Incremental EM is applied again to generate the final summary. This step has converted the multi-document summarization into single-document summarization. It can be well understood by the Figure 2.

Thus second tier summarization reduces the complexity of multi-document summarization by replacing it with single-document summarization. In our architecture we are considering Incremental EM for PLSA learning instead of EM. Incremental EM differs from EM only in E-step. In E- step of EM calculation of posterior variable $p(z_k|d_{i,} w_j)$ is done by using all values of the sets $p(w_j|z_{k,})$ and $P(z_k|s_i)$ whereas Incremental EM splits the set $P(z_k|s_i)$ into z subsets where z is the no of topics and updates the value of $p(z_k|d_{i,} w_j)$ with each subset of $P(z_k|s_i)$ and the set $p(w_j|z_{k,})$. This reduces the learning time of PLSA significantly and minimizes the summary generation time as well.

The main goal of our work is to produce a summary with minimized redundancy and maximized frequency of important sentences. The summary generation at tier-1, aims to reduce redundancy whereas the tier-2 summarization enhances the chances of more relevant information inclusion in the summary.

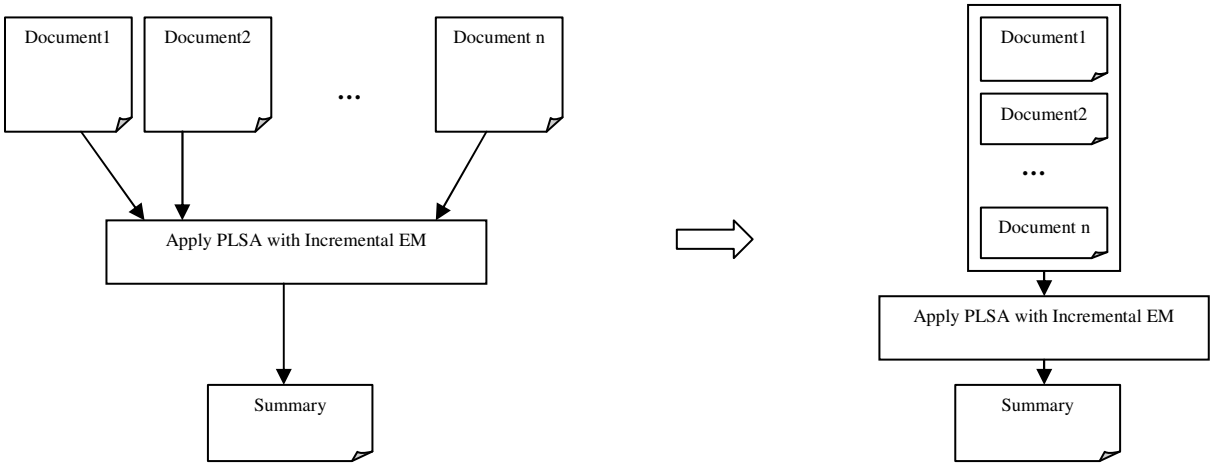

Figure 2. Multiple-document Summarization

Our research work takes the advantage of Single as well as Multi-document summarization both which results into more accurate and more relevant summary. This fact can be easily justified with our experimental results.

V. RESULTS AND DISCUSSION

We implemented and analyzed the performance of our proposal using java application on a 2GB memory Pentium machine. We also performed a brief analysis of the delay reduction in summarizer's response which was result of application of Incremental EM for PLSA learning. Table-1 shows the reduced delay time with respect to change in number of topics and the line graph in Figure 3 justifies the fact that with increase in number of topics Incremental EM enhances the performance improvement of PLSA. From the Table-1 it is clearly visible that with the increment in the number of topics the time consumption rate of Incremental EM is very slow.

Table 1. Performance Result of PLSA with Incremental EM

Number of topics (*z*)	**Reduced delay (*in seconds*)**
2	9
3	20
4	44
5	65
6	116
7	165

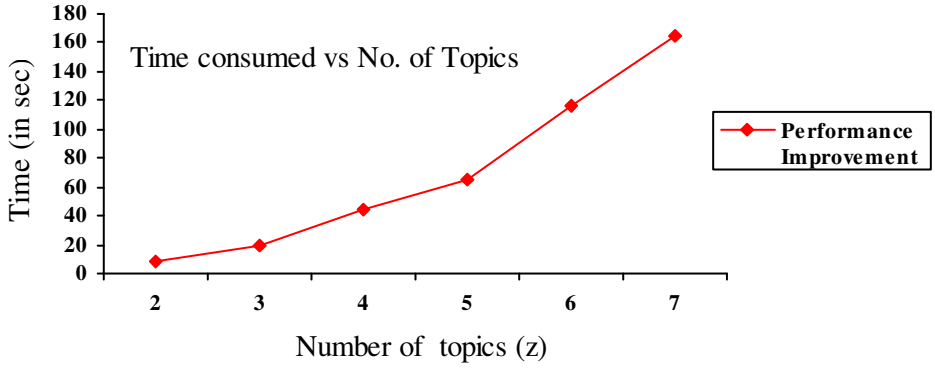

Figure 3. Performance Improvement graph of PLSA with Incremental EM

All generated summaries had been evaluated using ROUGE (Recall-oriented understudy for gisting Evaluation). It is a standard evaluation technique used by DUC. It contains measures to judge the

summary quality by comparing it to human created summaries [11]. For validation, we used ROUGE-L measure. The tabular representation of the resultant figures is given in Table-2. From Table-2 it is clearly visible that results obtained at tier-2 are better than tier-1. This is because tier-2 summary contains more relevant information.

Table 2. Summary evaluation results as ROUGE-L score

	Number of topics (z)	ROUGE-L
Tier-1	2	.573
	3	.562
	4	.534
	5	.510
Tier-2	2	.621
	3	.620
	4	.589
	5	.552

In this paper we argued that using Incremental EM for two tier summarization improves quality of processing. This improvement is achieved at the expense of some compression loss. But we found the compression loss is insignificant in comparison to improved processing time.

VI. CONCLUSIONS

In this paper we introduced two-tier architecture for summarization that is based on Probabilistic Latent Semantic Analysis with incremental Expectation Maximization algorithm. We presented every sentence as a distribution over latent topics and trained PLSA. Sentences are picked for summary based on their ranking. This process is performed at two levels to refine summary quality further. The results obtained from the experiments prove that resultant summary is of high quality and summarizer's processing time improvement is also very impressive. Tier-2 results outperform the tier-1 results for ROUGE-L score and shows that the final summary contains more relevant information and less redundancy. Another benefit of our work is that final summary contains sentences from all topics of the documents and we can apply this summarization architecture for documents written in any language. Hence our architecture provides a solid framework for summary generation. In future we will extend our work by time-stamping the dataset to generate generic solutions for opinion mining.

ACKNOWLEDGEMENTS

The authors would like to thank the Institution, Amity School of Engineering and Technology (ASET) with soul gratitude where the whole research is carried out.

REFERENCES

[1] H. Bhandari, M. Shimbo, T. Ito, and Y. Matsumoto, (2008), "Generic text summarization using probabilistic latent semantic indexing", *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp.133-140.

[2] Leonhard Hennig, (2009),"Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis". *Proceedings of International conference on Recent Advances in NLP*, pp. 144-149.

[3] J. xu, G. ye, Y. wang, G. Herman, B. Zhang, Jun Yang, (2009), "Incremental EM for Probabilistic Latent Semantic Analysis on Human Action Recognition", *Proceedings of Advanced Video and Signal Based Surveillance(AVSS'09),* Sixth IEEE international conference, pp.55-60.

[4] T. hofmann, (1999), "Probabilistic Latent Semantic Indexing", In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval,*, pp. 50-57.

[5] T. hofmann, (2001), "Unsupervised Learning by Probabilistic Latent Semantic Analysis", *Machine Learning,* Vol. 42,No.1-2, pp. 177-196.

[6] V. Gupta, G.S. Lehal, (2010), "A Survey of Text Summarization Extractive Techniques" ,*Journal of Emerging Technologies in Web Intelligence*, Vol. 2,No. 3, pp. 258-268.
[7] F.L. Wang, C. C. Yang and X. Shi, (2006), "Multi-document Summarization for Terrorism Information Extraction", *Springer-Verlag Berlin Heidelberg,* ISI 2006, LNCS 3975, pp. 602-608.
[8] Z. Zhuang, G. Karypis, X. Ning, Q. He, Z. Shi, (2012), "Multi-view learning via probabilistic latent semantic analysis", *Elsevier Information Sciences journal*, in press.
[9] Z. Sun, E. Lim, K. Chang, T.K. Ong and R. K. Gunaratna, (2005) ,"Event Driven Document Selection For Terrorism Information", *Springer-Verlag Berlin Heidelberg* ,ISI 2005,LNCS 3495, pp. 37-48.
[10] K. Zhou, G. R. Xue, Q. Yang and Y. Yu, (2010), "Learning with Positive and Unlabeled Examples Using Topic-Sensitive PLSA", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 1, pp 46-48.
[11] Chin-Yew Lin, (2004), "ROUGE: A Package for Automatic Evaluation of Summaries", *Proceedings of Workshop on Text Summarization*, post conference workshop of ACL.
[12] R. Barzilay, N. Elhadad, and K. McKeown, (2002), "Inferring strategies for sentence ordering in multi-document news summarization", *Journal of Artificial Intelligence Research*, 17:3, pp 35-55.
[13] C. Shen, T. Li, and C. Ding, (2011), "Integrating Clustering and Multi-Document Summarization by Bi-mixture Probabilistic Latent Semantic Analysis (PLSA) with Sentence Bases", *Proceedings of the twenty-fifth AAAI conference on Artificial intelligence* (AAAI-11).

Authors

Madhuri Singh is pursuing M.Tech in Computer Science and Engineering from Amity School of Engineering & Technology (ASET) at Amity University, Noida, Uttar Pradesh, India. She has done B.Tech in Information Technology from IET Faizabad in 2007. She has an excellent academic record and is a consistent performer. During her M.Tech she received scholarship for securing top rank. She is a member of International Association of Engineers (IAENG). She has also worked as a software trainee. She has implemented research projects such as credit card fraud detection using Hidden Markov Model etc. Her research interests include Information Retrieval and Data Mining.

Farhat Ullah Khan has done his M.Tech in Information Technology with specialization in Intelligent Systems, from Indian Institute of Information Technology Allahabad (IIITA) in the year 2010. He has served as a software developer in an IT company and worked as a freelancer web designer and developer during his MCA. He has also done BCA and PGDCA. He also has qualified Microsoft Certification (MCP) in ASP.Net using C#. He is a member of IEEE and IET UK. Currently he is an Assistant Professor in Computer Science and Engineering Dept. in Amity School of Engineering and Technology (ASET), at Amity University, Noida, Uttar Pradesh, India. Professor Khan is contributing in the research areas like Intelligent Systems, Natural Language Processing; Machine learning and soft computing techniques and applications. He is actively involved both in research and academia. He has implemented various research projects like Focused Crawler for Opinion Mining, Automatic Text Summary generator, Speech to text and text to speech converters etc. Real time voice processing, Finding Useful information on domain specific crawled data to establish a trend is among his current areas of research.