Effect of Incremental EM on Document Summarization using Probabilistic Latent Semantic Analysis

Madhuri Singh *Member IAENG*, Farhat Ullah Khan

Abstract— In our research work we have developed a summarizer that produces an effective and compact summary using probabilistic approach of LSA. Here we have used incremental EM instead of standard EM. In this paper we have shown, how the incremental EM affects summarization process. We have also performed a performance comparison experiment on the standard and incremental EM. It is concluded with preliminary experimental results that Standard EM requires large no of iterations for convergence in comparison to Incremental EM and makes PLSA training fast. In case of real time text summarization, time is the key. Thus speedy incremental EM enhances the pace of PLSA summarizer without losing the document integrity or summary quality. These summaries are generated on a dataset collected from various news portals. The results prove that incremental EM makes summarizer fast in comparison to standard EM.

Index Terms— EM, information retrieval, PLSA, Posterior probability, stemming

I. Introduction

FOR instant decision making, right information at right time is very crucial. Although internet is a very popular source of information but the information is scattered all over the web in a large document collection. It is very difficult to find required relevant information instantly. This problem can be reduced by document summarization. Summarization is the process of compressing a document and still preserving the overall sense [6]. Although various models have been introduced for document summarization but this research area is still left with lot of improvement possibilities. We have implemented a summarizer that minimizes the processing time to generate the summary. We have emphasized on single document extractive summary generation. Summaries can be generated in two ways- Extractive and Abstractive [6]. Generating abstractive summary is very complex process because it requires a brief understanding of the source document and technologies to create such summaries and these are still evolving. Extractive summaries are comparatively easy to generate as it works by simply selecting important sentences and paragraphs from the original document.

The methodology that we have used for the summarization is PLSA. PLSA distributes the document over latent variables and is based on the likelihood principal [4]. It represents the sentences as probability distributions over latent topics and considers a document as a mixture of topics. There are various other methods of summarization which are MMR (Maximal Marginal Relevance), graph based summarization using Lexpage rank & HITS and LSA [1]. The main disadvantage of these approaches (except LSA) is the resultant summary covers only a single topic of the document. Although LSA overcomes this disadvantage but it is not very popular because of its weak statistical foundation. Whereas PLSA eliminates this problem by treating each document as a collection of topics and resultant summary covers all important topics. PLSA is more reliable because of its strong statistical establishment.

The summarization process involves preprocessing of the original document, PLSA training, sentence score generation and finally summary generation. Preprocessing contains sentence splitting, stop word removal and stemming. For stemming we have used Porter's stemming algorithm. Once source document is preprocessed thereafter PLSA training starts. It uses EM algorithm to calculate and update various posterior probabilities. The main problem with it is that it is very time consuming which makes the summarization process slow. In our experiments we have found that the Incremental EM reduces the training time significantly hence enhances the summarizer's speed.

The rest of the paper is organized as follows. Section II briefly covers the previous related work. A brief Introduction of the PLSA is given in section III. Section IV explains our proposed work. Sections V, VI and VII cover experimental results, conclusion & future work respectively.

II. Related Work

To avoid the information overflow various summarization approaches have been developed over the years. Automatic document summarization [6] started in the early fifties. The early techniques used standard keyword, cue, title and location methods .The methods in nineties employed HMM and clustering.

Z. Sun and Lim [7] proposed an event driven document selection method that considers only domain specific documents and converts the various events in a document in the form of entities and relationships. Then various pattern

Manuscript received March 18, 2012; revised April 05, 2012.

Madhuri Singh is with the Amity School of Engineering and Technology, Amity University, Noida, U.P, India pursuing M.Tech in Computer Science & Engineering. IAENG Member number: 118837. Phone number: 091-9650001659; e-mail: immadhurisingh@ gmail.com.

Farhat Ullah Khan has provided guidance for this research work. He is now with the Department of Computer Science & Engineering, Amity University, Noida, U.P, India (e-mail: fukhan@amity.edu).

ISBN: 978-988-19252-1-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

based selection strategies are used to maximize information gain. In the work proposed by F. L. Wang [8], the various news stories are arranged in a hierarchical structure and then Fractal summarization model is used to generate summary. In 2008 [1] a PLSA summarizer for single document extractive summarization has been developed. It uses standard EM 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. It shows that PLSA is more suitable for capturing sparse information in a sentence than the LSI.

III. PLSA WITH EM

PLSA is probabilistic topical approach that considers the fact 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- Set of documents where

$d \in D = \{d_1, d_2, \ldots, d_j\}$

W- Set of words where

$w \in W = \{w_1, w_2, \ldots, w_i\}$

Z- 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

$$p(d,w) = p(d)p(w|d) \text{ where} \qquad \ldots(1a)$$

$$p(w|d) = \sum_{z \in Z} p(w|z)p(z|d) \qquad \ldots(1b)$$

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

$$l = \sum_{d \in D} \sum_{w \in W} n(d,w) \log P(d|,w) \qquad \ldots(1c)$$

To maximize (1a) in the presence of latent variables Expectation Maximization (EM) algorithm is used [4].

In case of single document summarization d represents a sentence s, z represents a topic and w represents a word [1]. The two steps of EM algorithm [5] is given below-

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

$$p(z_k|d_i, w_j) \ \alpha \ p(w_j|z_k)\, P(z_k|s_i) \qquad \ldots(2a)$$

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

$$P(z_k|d_i) \ \alpha \ \sum_j n(d_i, w_j) \, p(z_k|d_i, w_j) \qquad \ldots(2b)$$

$$P(w_j|z_k) \ \alpha \ \sum_i n(d_i, w_j) \, p(z_k|d_i, w_j) \qquad \ldots(2c)$$

$$P(z_k) \ \alpha \ \sum_{i,j} n(d_i, w_j) \, p(z_k|d_i, w_j) \qquad \ldots(2d)$$

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.

IV. PROPOSED WORK

In our work we have used incremental EM. Incremental EM is proved to be useful in case of application of PLSA on human action recognition [3]. Summarization starts with preprocessing. Once documents are preprocessed PLSA training takes place. Training starts with random values but with the application of incremental EM on each iteration training data becomes more accurate. Once PLSA is trained a score is generated for each sentence using following formula-

$$\text{sentence score} = \sum_k p(d_i|z_k) \, p(z_k) \qquad \ldots (3a)$$

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

$$\text{threshold} = ds + ds/nos \qquad \ldots (4a)$$

Where nos is the total number of sentences in the document. All sentences whose score is greater than the threshold are included in the summary. The main aim of our work is to maximize the processing speed and minimize the processing latency.

Here is a diagrammatic representation of basic work flow of our approach-

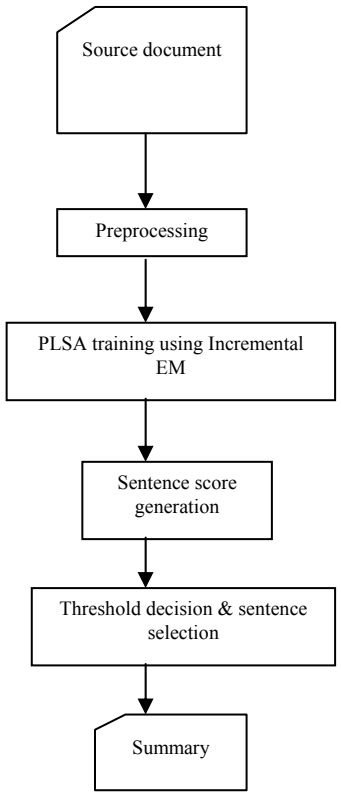

Fig. 1 Text Summarization process using Incremental EM

ISBN: 978-988-19252-1-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Our proposed algorithm for summarization is given below.

Algorithm

//The algorithm takes a source document as input and returns the summary as output
// s_i: a sentence where each $s \epsilon S$
//w_j: a word where each $w \epsilon W$
//z_k: a topic where each $z \epsilon Z$
//**n**: total number of subsets
//*sn*: total number of sentences in each subset
//***noz***: total number of topics
//***nos***: total number of sentences in the document
//***now***: total number of words in the document
//***M***: $p(z_k|s_i)$ a *noz*-by-*nos* matrix
//***N***: $p(w_j|z_k)$ a *now*-by-*noz* matrix
//***P***: $p(z)$ a *1*-by-*noz* matrix
//***Q***: $p(z_k|s_i, w_j)$ a *noz*-by-*nos*-by-*now* matrix
//***count***: increment variable
//***i***: integer variable
//***Str***: list of strings

1. Perform sentence splitting
2. Remove all stop words from the document
3. Perform Stemming
4. Start training PLSA and initialize *M,N* and *P* randomly
5. Calculate *Q* using 2a
6. Initialize *count*=0,*i*=0
7. Break the matrix *M* into *z* sub matrices M_i with dimension *z*-by-*sn*
8. While *i*<*z*
9. Start E-step and set
10. *count*= (*count*+1)mod(*nos*/*sn*)
11. Select sub matrix M_i
12. Calculate *Q* using M_i and *N* from 2a
13. Start M-step and update *M,N* and *P* using 2b,2c and 2d
14. If *i*<*z* then
15. go to 8
16. Else
17. End loop
18. End If
19. For each s_i calculate sentence score using *P* from 3a
20. End loop
21. Calculate threshold using 4a

//Picking sentences for the summary

22. For each s_i If *sentence score* > *threshold* then
23. add sentence in the *Str*
24. End loop
25. Return *Str*

Fig. 2 Pseudo code for summarization using PLSA with incremental EM

V. Results

The experiment was performed on the java application to test the effectiveness of our algorithm. The test was performed by varying no of topics *z* from 2 to 7. We also performed summarization using EM.

The results show that increment in the number of topics boosts the time consumed by EM sharply, which results in high processing delay and downgrades the summarizer's performance. Whereas, it has very less impact on processing time of Incremental EM. The line graph in Fig. 2 establishes and justifies the above mentioned fact. For large values of *z*, the summarization speed improvement with Incremental EM is very impressive. Hence for real time summarization applications PLSA with incremental EM is an excellent approach to work with.

TABLE I
Experimental Results

Number of topics (z)	Threshold value	Compression achieved (%)	Summarization time using PLSA with EM (*in seconds*)	Threshold value	Compression achieved (%)	Summarization time using PLSA with Incremental EM (*in seconds*)
2	1.79E-60	30.00	22	4.56E-60	30.00	13
3	8.20E-14	39.26	41	6.88E-10	36.60	21
4	1.66E-21	45.07	66	7.74E-15	34.50	22
5	1.88E-29	45.07	95	1.27E-18	34.50	30
6	1.34E-37	45.07	151	1.42E-23	32.74	35
7	1.34E-37	48.67	205	3.26E-28	31.00	40

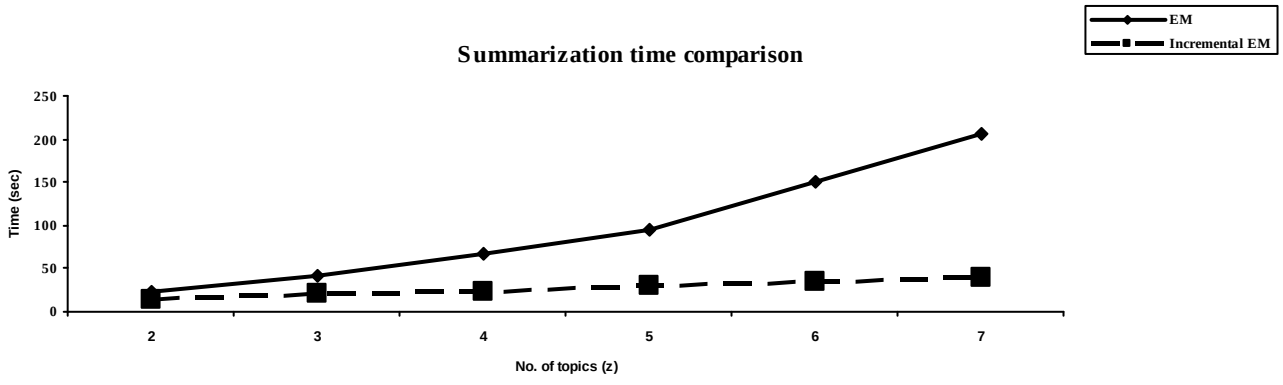

Fig. 3 Performance comparison graph of EM and incremental EM

VI. Conclusion

In this paper we have presented an algorithm for summarization that uses incremental EM. It reduces the summarization time significantly by updating the estimation parameters in E-step, using only a subset of teaching data. Experimental results prove that reduced convergence time of our algorithm speed ups the summarization process without losing significant compression.

Acknowledgment

Madhuri Singh thanks 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, "Generic text summarization using probabilistic latent semantic indexing", *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, 2008, pp.133-140.

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

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

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

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

[6] V. Gupta, G.S. Lehal.A, "Survey of Text Summarization Extractive Techniques" *„Journal of Emerging Technologies in Web Intelligence*, Vol. 2, No. 3, pp. 258-268, 2010.

[7] Z. Sun, E. Lim, K. Chang, T.K. Ong and R. K. Gunaratna, "Event Driven Document Selection For Terrorism Information", *Springer-Verlag Berlin Heidelberg* , ISI 2005, LNCS 3495, 2005, pp. 37-48.

[8] F.L. Wang, C. C. Yang and X. Shi, "Multi-document Summarization for Terrorism Information Extraction", *Springer-Verlag Berlin Heidelberg,* ISI 2006, LNCS 3975, 2006, pp. 602-608.

ISBN: 978-988-19252-1-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)