# Generic Text Summarization Using Probabilistic Latent Semantic Indexing

**Harendra Bhandari**
Graduate School of Information Science
Nara Institute of Science and Technology
Nara 630-0192, Japan
harendra-b@is.naist.jp

**Masashi Shimbo**
Graduate School of Information Science
Nara Institute of Science and Technology
Nara 630-0192, Japan
shimbo@is.naist.jp

**Takahiko Ito**
Graduate School of Information Science
Nara Institute of Science and Technology
Nara 630-0192, Japan
takahahi-i@is.naist.jp

**Yuji Matsumoto**
Graduate School of Information Science
Nara Institute of Science and Technology
Nara 630-0192, Japan
matsu@is.naist.jp

## Abstract

This paper presents a strategy to generate generic summary of documents using Probabilistic Latent Semantic Indexing. Generally a document contains several topics rather than a single one. Summaries created by human beings tend to cover several topics to give the readers an overall idea about the original document. Hence we can expect that a summary containing sentences from better part of the topic spectrum should make a better summary. PLSI has proven to be an effective method in topic detection. In this paper we present a method for creating extractive summary of the document by using PLSI to analyze the features of document such as term frequency and graph structure. We also show our results, which was evaluated using ROUGE, and compare the results with other techniques, proposed in the past.

## 1 Introduction

The advent of the Internet has made a wealth of textual data available to everyone. Finding a specific piece of information in this mass of data can be compared with "finding a small needle in a large heap of straw." Search engines do a remarkable job in providing a subset of the original data set which is generally a lot smaller than the original pile of data. However the subset provided by the search engines is still substantial in size. Users need to manually scan through all the information contained in the list of results provided by the search engines until the desired information is found. This makes automatic summarization the task of great importance as the users can then just read the summaries and obtain an overview of the document, hence saving a lot of time during the process.

Several methods have been proposed in the field of automatic text summarization. In general two approaches have been taken, extract-based summarization and abstract-based summarization. While extract-based summarization focuses in finding relevant sentences from the original document and using the exact sentences as a summary, abstract-based summaries may contain the words or phrases not present in the original document (Mani, 1999). The summarization task can also be classified as query-oriented or generic. The query-oriented summary presents text that contains information relevant to the given query, and the generic summarization method presents the summary that gives overall sense of the document (Goldstein et al, 1998). In this paper, we will focus on extract-based generic single-document summarization.

In the recent years graph based techinques have become very popular in automatic text summariza-

tion (Erkan and Radev, 2004), (Mihalcea, 2005). These techniques view each sentence as a node of a graph and the similarities between each sentences as the links between those sentences. Generally the links are retained only if the similarity values between the sentences exceed a pre-determined threshold value; the links are discarded otherwise. The sentences are then ranked using some graph ranking algorithms such as HITS (Kleinberg, 1998) or PageRank (Brin and Page, 1998) etc. However the graph ranking algorithms tend to give the highest ranking to the sentences related to one central topic in the document. So if a document contains several topics, these algorithms will only choose one central topic and rank the sentences related to those topic higher than any other topics, ignoring the importance of other topics present. This will create summaries that may not cover the overall topics of the document and hence cannot be considered generic enough. We will focus on that problem and present a way to create better generic summary of the document using PLSI (Hofmann 1999) which covers several topics in the document and is closer to the summaries created by human beings. The benchmarking done using DUC[2] 2002 data set showed that our technique improves over other proposed methods in terms of ROUGE[1] evaluation score.

## 2 Related Work

### 2.1 Maximal Marginal Relevance(MMR)

MMR is a summarization procedure based on vector-space model and is suited to generic summarization (Goldstein et al, 1999). In MMR the sentence are chosen according to the weighed combination of their general relevance in the document and their redundancy with the sentences already chosen. Both the relevance and redundancy are measured using cosine similarity. Relevance is the cosine similarity of a sentence with rest of the sentence in the document whereas redundancy is measured using cosine similarity between the sentence and the sentences already chosen for the summary.

### 2.2 Graph Based Summarization

The graph-based summarization procedure are be coming increasingly popular in recent years. Lex PageRank (Erkan and Radev, 2004) is one of such methods. LexPageRank constructs a graph where each sentence is a node and links are the similarities between the sentences. Similarity is measured using cosine similarity of the word vectors, and if the similarity value is more than certain threshold value the link is kept otherwise the links are removed. PageRank is an algorithm which has been successfully applied by Google search engine to rank the search results. Similarly PageRank is applied in LexPageRank to rank the nodes (or, sentences) of the resultant graph. A similar summarization method has been proposed by Mihalcea (2005).

Algorithms like HITS and PageRank calculate the principal eigenvector (hence find the principal community) of the matrix representing the graph. But as illustrated in Figure 1, another eigenvector which is slightly smaller than the principal eigenvector may exist. In documents, each community represented by the eigenvectors can be considered as a topic present in the document. As these algorithms tend to ignore the influence of eigenvectors other than largest one, the sentences related to topics other than a central one can be ignored, and creating the possibility for the inclusion of redundant sentences as well. This kind of summary cannot be considered as a generic one.

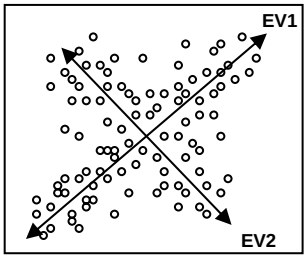


Figure 1. In algorithms like HITS and PageRank only the principal eigenvectors are considered. In the figure the vector EV1 is slightly larger than vector EV2, but the score commanded by members of EV2 communities are ignored.

As we mentioned in section 1, we take into consideration the sentences from all the topics generated by PLSI in the summary, hence getting a more generic summary.

### 2.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) (Deerwester et al.,

[1] ROUGE:http://openrouge.com/default.aspx
[2] http://duc.nist.gov

1990) takes the high dimensional vector space representation of the document based on term frequency and projects it to lesser dimension space. It is thought that the similarities between the documents can be more reliably estimated in the reduced latent space representation than original representation. LSA has been applied in areas of text retrieval (Deerwester et al., 1990) and automatic text summarization (Gong and Liu, 2001). LSA is based on Singular Value Decomposition (SVD) of $m \times n$ term-document matrix A. Each entry in A, $A_{ij}$, represents the frequency of term $i$ in document $j$. Using SVD, the matrix A is decomposed into U,S,V as,

$A=USV^T$

U=Matrix of $n$ left singular vectors

S=diag($\sigma_i$)=Diagonal matrix of singular values where with $\sigma_i \geq \sigma_{i+1}$ for all $i$.

$V^T$=Matrix of right singular vectors. Each row represents a topic and the values in each row represent the score of documents, represented by each columns, for the topic represented by the row.

Gong and Liu (2001) have proposed a scheme for automatic text summarization using LSA. Their algorithm can be stated below.

a. Choose the highest ranked sentence from $k^{th}$ right singular vector in matrix $V^T$ and use the sentence in summary.
b. If $k$ reaches the predefined number, terminate the process; otherwise, go to step a again.

LSA categorizes sentences on the basis of the topics they belong to. Gong and Liu's method picks sentences from various topics hence producing the summaries that are generic in nature.

In section 3 we explain how PLSI is more advanced form of LSA. In section 5, we compare our summarization results with that of LSA.

## 3 Probabilistic Latent Semantic Indexing

Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999) is a new approach to automated document indexing, and is based on a statistical latent class model for factor analysis of count data. PLSI is considered to be a probabilistic analogue of Latent Semantic Indexing (LSI), which is a document indexing technique based on LSA. Despite the success of LSI, it is not devoid of deficits. The main argument against LSI is pointed to its unsatisfactory statistical foundations. In contrast, PLSI has solid statistical foundations, as it is based on the maximum likelihood principle and defines a proper generative model of data. Hofmann (1999) has shown that PLSI indeed performs better than LSI in several text retrieval experiments. The factor representation obtained in PLSI allows us to classify sentences according to the topics they belong to. We will use this ability of PLSI to generate summary of document that are more generic in nature by picking sentences from different topics.

## 4 Summarization with PLSI

### 4.1 The Latent Variable Model for Document

Our document model is similar to Aspect Model (Hofmann et al, 1999, Saul and Pereira, 1997) used by Hoffman (1999). The model attempts to associate an unobserved class variable $z \in Z=\{z_1, ..., z_k\}$ (in our case the topics contained in the document), with two sets of observables, documents $(d \in D=\{d_1, .....d_m\}$, sentences in our case) and words $(w \in W=\{w_1, \ldots, w_n\})$ contained in documents. In terms of generative model it can be defined as follows:

-A document $d$ is selected with probability $P(d)$

-A latent class $z$ is selected with probability $P(z|d)$

-A word $w$ is selected with probability $P(w|z)$

For each document-word pair $(d,w)$, the likelihood for each pair can be represented as

$$P(d,w)=P(d)P(w|d)=P(d)\sum_z P(w|z)P(z|d).$$

Following the maximum likelihood principle $P(d)$, $P(z|d)$, $P(w|z)$ are determined by the maximization of of log-likelihood function,

$$L=\sum_d \sum_w n(d,w)\log P(d,w)$$

where $n(d,w)$ denotes the term frequency, i.e., the number of time $w$ occurred in $d$.

### 4.2 Maximizing Model Likelihood

Expectation Maximization (EM) is the standard procedure for maximizing likelihood estimation in the presence of latent variables. EM is an iterative procedure and each of the iteration contains two steps. (a) An Expectation (E) step, where the posterior probabilities for latent variable $z$ are computed and (b) Maximization (M) step, where parameters for given posterior probabilities are computed.

The aspect model can be re-parameterized using the Bayes' rule as follows:

$$P(d,w)= \sum_z P(z)\, P(d|z)\, P(w|z)\ .$$

Then using the re-parameterized equation the E-step calculates the posterior for $z$ by

$$P(z \mid d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')}$$

This step calculates the probability that word $w$ present in document $d$ can be described by the factor corresponding to $z$. Subsequently, the M-step re-evaluates the parameters using following equations.

$$P(w \mid z) = \frac{\sum_d n(d,w)P(z|d,w)}{\sum_{d,w'} n(d,w')P(z|d,w')}, \quad (1)$$

$$P(d \mid z) = \frac{\sum_w n(d,w)P(z|d,w)}{\sum_{d',w} n(d',w)P(z|d',w)}, \quad (2)$$

$$P(z) = \frac{\sum_{d,w} n(d,w)P(z|d,w)}{\sum_{d,w} n(d,w)} \quad (3)$$

Alternating the E- and M- steps one approaches a converging point which describes local maximum of the log-likelihood.

We used the tempered EM (TEM) as described by Hofmann (1999). TEM basically introduces a control parameter $B$, upon which the E-step is modified as,

$$P(z \mid d, w) = \frac{P(z)[P(d|z)P(w|z)]^B}{\sum_{z'} P(z')[P(d|z')P(w|z')]^B} \quad (4)$$

The TEM reduces to original EM if $B$=1.

### 4.3 Summarization procedure

We applied PLSI in 4 different ways during the summarization process. We will denote each of the 4 ways as **PROC1**, **PROC2**, **PROC3**, **PROC4**. Each of the four summarization procedure is discussed below.

**PROC1 (Dominant topic only)**: PROC1 consists of the following steps:

a. Each document is represented as term-frequency matrix.
b. *P(w|z), P(d|z), and P(z)* (as in (1), (2), (3)) are calculated until the convergence criteria for EM-algorithm is met. *P(d|z)* represents the importance of document d in given topic represented by *z* and *P(z)* represents the importance of the topic z itself in the document d.
c. *z* with highest probability *P(z)* is picked as the central topic of the document and then the sentences with highest *P(d|z)* score contained in selected topic are picked.
d. The top scoring sentences are used in the summary.

**PROC2 (Dominant topic only)**: PROC2 is the graph based method. PROC2 is similar to PROC1 except for the fact that instead of using term-frequency matrix we use sentence-similarity matrix. Sentence-similarity matrix A is $n \times n$ matrix where $n$ is the number of sentences present in the document. Cosine similarity of each sentence present in the document with respect to all the sentences is calculated. The cosine-similarity values calculated are used instead of term-frequency values as in PROC1. Each entry $A_{ij}$ in matrix A is 0 if the cosine similarity value between sentence $i$ and sentence $j$ is less than threshold value and 1 if greater. We used 0.2 as the threshold value in our experiments after normalizing cosine similarity value. Steps b, c, d from PROC1 are followed after the initial procedure is complete.

This method is analogous to PHITS (Cohn and Chang (2001)) method where the authors utilized PLSI to find communities in hyperlinked environment.

**PROC3 (Multiple topics)**: In both PROC1 and PROC2 we did not take the advantage of the fact that PLSI divides a document into several topics. We only used the sentences from highest ranked topic. In PROC3 we attempt to combine the sentences from different topics while forming the summary. PROC3 can be explained in the following steps.

a. Steps a and b from PROC1 are taken as normal.
b. We mentioned that *P(d|z)* represents the score of the sentence *d* in topic *z*. In this procedure we will create new score R for each sentence using following relation.

$$R=\sum_z P(d|z)P(z)=P(d)$$

Table 1: Evaluation of summaries

The table shows the score of summaries generated using methods described in section 4.3. On the table n means number of topics into which the document has been divided into. Control parameter B from (4) was fixed to 0.75 in this case.

| Method Used | **n** | ROUGE-L (recall) | Rouge1 | Rouge-2 | Rouge-SU4 |
|---|---|---|---|---|---|
| PROC1 | 2 | 0.499 | 0.557 | 0.242 | 0.272 |
| PROC2 | 2 | 0.465 | 0.515 | 0.227 | 0.253 |
| PROC3 | 2 | **0.571** | **0.634** | **0.291** | **0.321** |
| | 3 | 0.571 | 0.628 | 0.288 | 0.318 |
| | 4 | 0.571 | 0.62 | 0.28 | 0.31 |
| | 5 | 0.571 | 0.613 | 0.274 | 0.305 |
| | 6 | 0.5 | 0.612 | 0.27 | 0.302 |
| PROC4 | 2 | 0.473 | 0.508 | 0.225 | 0.25 |
| | 3 | 0.472 | 0.504 | 0.22 | 0.245 |
| | 4 | 0.472 | 0.5 | 0.219 | 0.244 |
| | 5 | 0.472 | 0.492 | 0.213 | 0.238 |
| | 6 | 0.471 | 0.483 | 0.207 | 0.231 |
| **Compared Methods** | | | | | |
| *LexPageRank | | 0.522 | 0.577 | 0.265 | 0.291 |
| *LSA | | 0.414 | 0.463 | 0.186 | 0.215 |
| *HITS | | 0.504 | 0.562 | 0.251 | 0.282 |

This will essentially score the sentences with generic values or the sentences which have good influence ranging over several topics better.

c. We pick the sentences that score highest score R as the summary.

PROC3 will pick sentences from several topics resulting in better generic summary of the document.

**PROC4 (Multiple Topics)**: PROC4 is essentially PROC3 except for the first few steps. PROC4 does not use the matrix created in PROC1 instead it uses the similarity-matrix produced in PROC2. Once the similarity matrix is created *P(z)* and *P(d|z)* are calculated as in step b of PROC1. Then steps b and c of PROC3 are taken to produce the summary of the document.

## 5 Experiments and Results

We produced summaries for all the procedures mentioned in section 4.3. We used DUC 2002 data set for summarization. DUC 2002 contains test data for both multiple document and single document summarization. It also contains summaries created by human beings for both single document and multiple document summarization. Our focus in this paper is single document summarization.

After creating summaries we evaluated summaries using ROUGE. ROUGE has been the standard benchmarking technique for summarization tasks adopted by Document Understanding Conference (DUC). We also compared our results with other summarization methods such as LexPageRank (Erkan and Radev, 2004) and Gong and Liu's (2001) LSA-based method. We also compared the results with HITS based method which is similar to LexPageRank but instead of PageRank, HITS is used as ranking algorithm (Klienberg 1998).The results are listed in Table 1.

We used five measures for evaluation, Rouge-L Rouge1, Rouge2, Rouge-SU4 and $F_1$. These methods are standard methods used in DUC evaluation

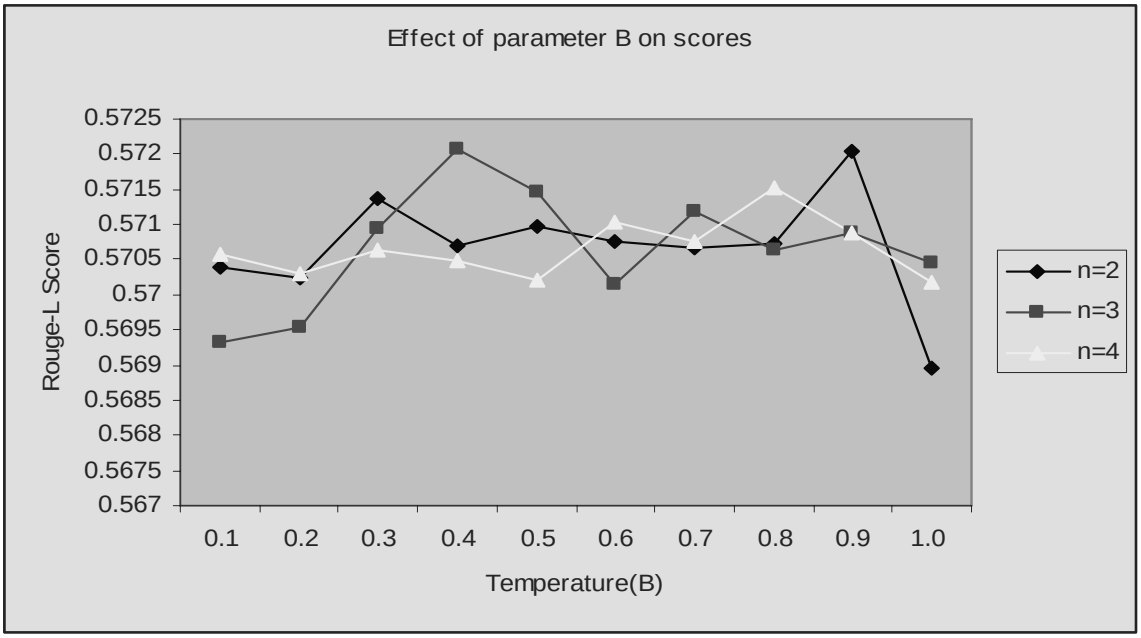


Figure 2: Effect of tempering factor B in the ROUGE-L score for PROC3.

tests and these schemes are known to be very effect tive to calculate the correlation between the summaries. All of the scores can be calculated using Rouge package. Rouge is based on N-gram statistics (Lin and Hovy, 2003). Rouge has been known to highly correlate with human evaluations. According to (Lin and Hovy, 2003), among the methods implemented in ROUGE, ROUGE-N (N=1,2), ROUGE-L, ROUGE-S are relatively simple and work very well even when the length of summary is quite short, which is mostly the case in single document summarization. ROUGE-N, ROUGE-L and ROUGE-S are all basically the recall scores. As DUC keeps the length of the summaries constant recall is the main evaluation criterion. F-measure is also shown in the table as a reference parameter, but since we kept the length of our summaries constant, too, the ROUGE-L, ROUGE-N and ROUGE-S scores carry the highest weight.

As seen on Table 1, the scores gained by PROC1 and PROC2 are less than others. This is mainly because the sentences chosen by these methods were simply chosen from one topic. As PROC3 and PROC4 use sentences from several topics the score of PROC3 and PROC4 were better than PROC1 and PROC2. For methods PROC3 and PROC4 we took the summaries for topics 2 through 6 and found that the method performed well when the number of topics was kept between 2 to 4. But the difference was very small, and in general the performance was quite stable.

We also compared our results to other methods such as LexPageRank and LSA and found that PROC3 performed quite well when compared to those methods. LexPageRank was marginally better in F-measure ($F_1$) but PROC3 got best recall scores. PROC3 also outperformed LSA by 0.16 in recall (ROUGE-L) scores. Comparison to HITS also shows PROC3 more advantageous.

## 6 Discussion

In this paper we have argued that choosing sentences from multiple topics makes a better generic summary. It is especially true if we compare our method to graph based ranking methods like HITS and PageRank. Richardson and Domingos (2002) have mentioned that both HITS and PageRank suffer from the topic drift. This not only makes these algorithms susceptible for exclusion of important sentences outside the main topic but miss the sentences from main topic as well. Cohn and Chang (2001) also have shown similar results for HITS. They (Cohn and Chang) have shown that the central topic identified by HITS (principal eigenvector) may not always correspond to the most authoritative topic. The main topic in fact may be represented by smaller eigenvectors rather than the principal one. They also show that the topic segre-

gation in HITS is quite extreme so if we just use principal eigenvector, first there is a chance of being drifted away from the main topic hence producing low quality summary and there is also a chance of missing out other important topics due to the extreme segregation of communities. In PLSI the segregation of topics is not as extreme. If a sentence is related to several topics the sentence can attain high rank in many topics.

We can see from the scores that the performance of graph based algorithms like LexPageRank and HITS are not as good as our method. This can be attributed to the fact that the graph based summarizers take only a central topic under consideration. The method that proved most successful in our summarization was the one where we extracted the sentences that had the most influence in the document.

We used the tempered version of EM-algorithm (4) in our summarization task. We evaluated the effect of tempering factor *B* in performance of summarization for PROC3. We found that that the tempering factor did not influence the results by a big margin. We conducted our experiment using values of **B** from 0.1 through 1.0 incrementing each step by 0.1. The results are shown in Figure 2. In the results shown in Table 1 the value for tempering factor was set to 0.75.

## 7 Conclusion and Future Work

In this paper we presented a method for creating generic summaries of the documents using PLSI. PLSI allowed us classify the sentences present in the document into several topics. Our summary included sentences from all the topics, which made the generation of generic summary possible. Our experiments showed that the results we obtained in summarization tasks were better than some other methods we compared with. LSA can also be used to summarize documents in similar manner by extracting sentences from several topics, but our experiments showed that PLSI performs better than LSA. In the future we plan to investigate how more recent methods such as LDA (Blei et al) perform in document summarization tasks. We also plan to apply our methods to multiple document summarization.


## 8 Acknowledgement

We pay our special gratitude to the reviewers who have taken their time to give very useful comments on our work. The comments were very useful for us to as we were able to provide wider perspective on our work with the help of those comments.


## References:


Blei D, Ng A, and Jordan M.2003. Journal of Machine Learning Research 3 993-1022.

Brin S and Page L.1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks 30(1-7): 107-117.

Carbonell J and Goldstein J.1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. Proc. ACM SIGIR

Cohn D, Chang H.2001. Learning to probabilistivally identify authoritative documents. Proceedings of 18th International Conference of Machine Learning.

Deerwester S, Dumais ST, Furnas GT, Landauer TK, and Harshman R.1990. Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science.

Erkan G and Radev DR.2004. LexPageRank: Prestige in Multi-Document Text Summarization.EMNLP.

Gong Y and Liu X.2001.Generic text Summarization using relevance measure and latent semantic analysis.Proc ACM SIGIR.

Hofmann, T.1999.Probabilistic Latent Semantic Indexing. Twenty Second International ACM-SIGIR Conference on Informatioln Retrieval.

Hofmann, et al.1998. Unsupervised Learning from Dyadic Data. Technical Report TR-98-042, International Computer Science Insitute, Berkeley, CA.

Kleinberg J.1998. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms.

Mani I. 1999. Advances in Automatic Text Summarization. MIT Press, Cambridge, MA, USA.

Mihalcea R.2005.Language Independent Extractive Summarization. AAAI

Richardson M, Domingos P.2002. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. Advances in Neural Information Processing Systems 14

Saul L and Pereria F.1997.Aggregate and mixed-order Markov models for statistical language processing.Proc 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval.