Extracting Topic Words and Clustering Documents by Probabilistic Graphical Models *

Hyung-Joo Shin and Byoung-Tak Zhang
Artificial Intelligence Lab (SCAI)
School of Computer Engineering, Seoul National University
Seoul, 151-742, Korea
{hjshin, btzhang}@scai.snu.ac.kr

ABSTRACT

We present a method for clustering documents and extracting topic words of each cluster using a probabilistic graphical model. We maximize the likelihood of the model with the Expectation Maximization algorithm. Our experiments demonstrate that the latent variables of the model can be seen as clusters of documents and terms.

1. INTRODUCTION

With the advent of the huge amount of text documents, the necessity of automatic clustering and extracting topic words of text documents has increased. In this paper, we present an algorithm for clustering documents and extracting topic words of each cluster using a probabilistic graphical model. We assume that if we can find out the hidden structures of a set of documents, we can classify the documents into clusters and extract topic words of each topic, that is, the clusters can be interpreted as topics. Hidden structure means the generating probabilities of words and documents given the probability of topics. This algorithm makes use of the Probabilistic Latent Semantic Indexing (PLSI), which has been shown to have strong results when applied to automatic indexing and information retrieval [3]. So we call this algorithm PLS_cluster from now.

2. LEARNING PROBABILISTIC GRAPHICAL MODELS FOR CLUSTERING AND EXTRACTING TOPIC WORDS

Aspect model used in PLSI assumes that observable variables, terms $w_j \in W = \{w_1, ..., w_M\}$ and documents $d_i \in D = \{d_1, ..., d_N\}$ are generated conditioned on the unobservable(latent) variables $z_k \in Z = \{z_1, \cdots, z_K\}$. The model uses count data (d_i, w_j), which is the frequency of the term w_j in the document d_i. This is a generative model whose likelihood to maximize is as follows:

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log P(d_i, w_j). \tag{1}$$

where $n(d_i, w_j)$ denotes the term w_j's frequency in the document d_i and

$$P(d_i, w_j) = \sum_{k=1}^{K} P(z_k) P(w_j|z_k) P(d_i|z_k). \tag{2}$$

*This research was supported by the Korea Science and Engineering Foundation (KOSEF) under grant 981-0920-107-2 and by the Korea Ministry of Information and Telecommunications under Grand C1-98-0068-00 through IITA.

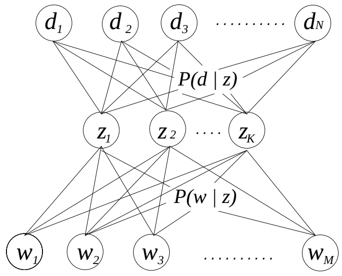

Figure 1: Probabilistic Graphical Model for Extracting Topic Words and Clutering Documents.

To maximize this likelihood, we fit $P(z)$, $P(d|z)$, and $P(w|z)$ with Expectation Maximization(EM) algorithm. E-step follows

$$P(z_k|d_i, w_j) = \frac{P(z_k)P(d_i|z_k)P(w_j|z_k)}{\sum_{k=1}^{K} P(z_k)P(d_i|z_k)P(w_j|z_k)}. \tag{3}$$

M-step follows:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{j=1}^{M} \sum_{i=1}^{N} n(d_i, w_j) P(z_k|d_i, w_j)} \tag{4}$$

$$P(d_i|z_k) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{j=1}^{M} \sum_{i=1}^{N} n(d_i, w_j) P(z_k|d_i, w_j)} \tag{5}$$

$$\begin{aligned} P(z_k) &= \tfrac{1}{R} \textstyle\sum_{j=1}^{M} \sum_{i=1}^{N} n(d_i, w_j) P(z_k|d_i, w_j), \\ R &\equiv \textstyle\sum_{j=1}^{M} \sum_{i=1}^{N} n(d_i, w_j). \end{aligned} \tag{6}$$

The detail of the equations is given in [3]. A simple sketch of this graphical model is shown in Fig 1. $P(w_j|z_k)$ and $P(d_i|z_k)$ can be interpreted as probabilities of the generating term w_j and document d_i from topic z_k. Let the model be $K = L$ where L denotes the known number of topics within the set of documents and K denotes the number of latent variables of the model. Then the topic of document d_i can be selected as follows:

$$topic'(d_i) = \arg\max_k P(d_i|z_k), k = 1, 2, 3, \cdots, K. \tag{7}$$

And the topic words for each topic can be selected as the words of high probabilities $P(w_j|z_k)$.

3. EXPERIMENTAL RESULTS

We have performed experiments with the subset of TREC8 adhoc collection. It includes documents of 50 topics. Human experts have decided which documents are relevant to each topic. Among 50 topics, we selected 4 topics (topic 401, 434, 439, 450) in which the relevant documents are of

Table 1: Topic Descriptions and most frequent words for topic 401, 434, 439, and 450.

Topics	Descriptions
Topic 401	Foreign minorities, Germany
	german, germani, year, foreign, mr, countri, state, govern, parti, minist, asylum, peopl, develop, nation, polit, report, percent, east, time, turki, wing, european, law, ...
Topic 434	Estonia, economy
	percent, bank, estonia, state, privat, year, enterprise, million, russian, foreign, compani, govern, trade, loan, countri, econom, polish, invest, product, fund, estonian, price, ...
Topic 439	inventions, scientific discoveries
	research, develop, techonologi, mar, materi, system, bridgeston, environ, compani, environment, process, industri, nuclear, basic, product, electr, high, wast, energi, amp, ...
Topic 450	King Hussein, peace
	jordan, peac, isreal, king, palestinian, jordanian, arab, isra, meet, al, state, talk, husayn, mr, crystallian, presi, process, majesti, issu, negoti, region, plo, time, countri, ...

Table 2: Extracted Topic Words using PLS_cluster (ordered according to $P(w|z_k)$).

Cluster #	Extracted Topic Words
$k = 2$	german, germani, mr, parti, year, foreign, peopl, countri, govern, asylum, polit, nation, law, minist, europ, state, immigr, democrat, social, turkish, west, east, attack, union,...
$k = 4$	percent, estonia, bank, state, privat, russian, year, enterprise, trade, million, trade, estonian, econom, countri, govern, compani, foreign, baltic, polish, loan, invest, fund, product,...
$k = 3$	research, techonologi, develop, mar, materi, system, nuclear, environment, electr, process, product, power, energi, control, japan, pollution, structur, chemic, plant,...
$k = 1$	jordan, peac, isreal, palestinian, king, isra, arab, meet, talk, husayn, agreem, presid, majesti, negoti, minist, visit, region, arafat, secur, peopl, east, washington, econom, sign, relat, jerusalem, rabin, syria, iraq,...

large number. We used 2216 distinct words with higher frequencies after stemming and stopwords elimination. We set $K = L = 4$.

Table 1 shows topic descriptions and the terms whose frequencies are high in each topic. Table 2 shows the extracted topic words (most probable words in each cluster) of 4 latent variables using PLS_cluster. This shows that the model can extract topic words without pre-labeled training sets.

Table 3, 4, and 5 are the confusion matrixes for TREC8 data using PLS_cluster, naive Bayesian classification [2] (supervised learning), Self-Organizing Maps [1] (unsupervised learning). Naive Bayes classifier is the simplest probabilistic graphical model which performs classification very well. PLS_cluster can cluster the documents into topics as well as naive Bayesian classifier, and better than another unsupervised learning algorithm, SOM in precision and recall.

4. CONCLUSIONS

In this paper, we presented an algorithm for clustering documents and extracting topic words of each cluster using a probabilistic graphical model used in PLSI. We could find the hidden structures (represented by density probability $P(w|z)$, $P(d|z)$) of the set of documents, and with this density probability we could classify documents into clusters which can be seen as topics. In addition, we could extract topic words which characterize each topic.

As we can see from the comparative results, supervised learning (here we used naive Bayesian Classification) is better at document classification, but requires a large amount of training data, which is not required for PLS_cluster presented in this paper. The clustering result using another popular clustering algorithm (SOM) is not as good as PLS_cluster in documents clustering.

Unfortunately, the current formulation of this model is computationally very expensive, compared to the cost of naive Bayesian classification or SOM clustering. And this model is not an adaptive one, not able to be fitted further with new-coming documents. More seriously, it requires that the number of latent variables (topics) should be known ahead of time. It would be more useful to use the model-based approach and some measure of fit to select the correct number of latent variables. We hope to improve this model by solving these problems.

Table 3: Confusion matrix using PLS_cluster.

	Label(Maximum)					
Topic(#doc)	k =2	4	3	1	Precision	Recall
401(300)	**279**	1	0	20	0.902	0.930
434(347)	20	**238**	10	79	0.996	0.686
439(219)	7	0	**203**	9	0.953	0.927
450(293)	3	0	0	**290**	0.729	0.990
	Label(Threshold: $P(d_i|z_k) \geq 0.001$)					
Topic(#doc)	k =2	4	3	1	Precision	Recall
401(300)	**279**	12	6	7	0.705	0.930
434(347)	60	**295**	20	49	0.883	0.850
439(219)	22	17	**210**	4	0.868	0.959
450(293)	35	10	6	**285**	0.826	0.973

Table 4: Confusion matrix using naive Bayesian classification (test set=0.6).

Topic(#doc)	k =1	2	3	4	Precision	Recall
401(185)	**171**	8	1	5	0.929	0.924
434(143)	0	**126**	0	17	0.783	0.881
439(73)	0	8	**65**	0	0.929	0.890
450(239)	13	19	4	**203**	0.902	0.849

Table 5: Confusion matrix using SOM.

Topic(#doc)	k =2	3	1	4	Precision	Recall
401(300)	**151**	28	100	21	0.696	0.503
434(347)	60	**199**	86	2	0.833	0.574
439(219)	4	12	**183**	20	0.470	0.836
450(293)	2	0	20	**271**	0.863	0.925

5. REFERENCES

[1] Kohonen, T., The Self-organizing map, in *Proceedings of IEEE*, vol. 78, pages 1464-1479, 1990.

[2] N. Friedman, D. Geiger, and M. Goldszmidt., Bayesian networks classifiers, in *Machine Learning*, vol. 29, pages 131-163, 1997.

[3] Thomas Hoffmann, Probabilistic Latent Semantic Indexing, in *Proceedings of the 22th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50-57, 1999.