Topic Extraction from Text Documents Using Multiple-Cause Networks

Jeong-Ho Chang[1], Jae Won Lee[2], Yuseop Kim[3], and Byoung-Tak Zhang[1]

[1] School of Computer Science and Engineering, Seoul National University, Seoul, Korea, 151-742
{jhchang, btzhang}@bi.snu.ac.kr
[2] School of Computer Science and Engineering, Sungshin Women's University, Seoul, Korea, 136-742
jwlee@cs.sungshin.ac.kr
[3] Ewha Institute of Science and Technology, Ewha Woman's University, Seoul, Korea, 120-750
yskim01@ewha.ac.kr

Abstract. This paper presents an approach to the topic extraction from text documents using probabilistic graphical models. Multiple-cause networks with latent variables are used and the Helmholtz machines are utilized to ease the learning and inference. The learning in this model is conducted in a purely data-driven way and does not require prespecified categories of the given documents. Topic words extraction experiments on the TDT-2 collection are presented. Especially, document clustering results on a subset of TREC-8 ad-hoc task data show the substantial reduction of the inference time without significant deterioration of performance.

1 Introduction

Due to the popularity of the Internet and the advance in digital library, we have seen an exponential increase in the amount of on-line text and digitized documents. But the abundance of large databases itself does not lead us to easy acquisition of relevant information, since it can be a tiresome and time-consuming work to look over and organize all documents about various topics. As a result, the demand for automatic text analysis – such as clustering, classification, summarization – is more increasing than ever before.

Recently, there have been much research in automatic topic extraction and semantic analysis, as a tool for data-driven text analysis, such as Latent Semantic Analysis (LSA) [3] using singular value decomposition. Similar approaches based on probabilistic generative models have been also proposed. These includes the Probabilistic Latent Semantic Analysis (PLSA) [7] and the Non-negative Matrix Factorization (NMF) [8], where documents are modelled by multinomial distribution and Poisson distribution, respectively. It has been shown that these approaches can enhance performance in such applications as language modelling, text categorization, and information retrieval.

M. Ishizuka and A. Sattar (Eds.): PRICAI 2002, LNAI 2417, pp. 434–443, 2002.

In this paper, we present an effective probabilistic approach based on the multiple-cause models [2][5][10] for topic extraction from unstructured text documents. A topic is represented by a set of its relevant or related words, and a document is assumed to be the result of a combination of one or more topics. If a specific topic is activated, it causes its related words more likely to appear and others less likely in a document. In the case that more than one topic are activated simultaneously, a document is generated by the cooperation of these topics.

To ease the learning and inference in the model, we adopt an approximation method based on the Helmholtz machine [1] and use an on-line stochastic algorithm for the model fitting, unlike the works in [9] where a deterministic gradient ascent method is used in the application of their multiple-cause mixture model to the text categorization. Additionally, we introduce some heuristics for incorporating the word frequency information in the binary-valued Helmholtz machines.

2 The Multiple-Cause Model

In the multiple-cause model, it is postulated that a data naturally arises from the consequences of a set of hidden causes [5]. Formally, the model can be represented by multiple-cause networks, and we describe these networks in terms of the analysis of text documents.

2.1 Multiple-Cause Networks

Figure 1 shows the typical form of the multiple-cause network [5]. Topics are represented by latent nodes and words in the vocabulary are by input nodes. Each topic z_k is a binary variable indicating its presence or absence, and is described as a set of independent *Bernoulli* distributions, one distribution for each term. When it is given a document collection $D = \{d_1, d_2, \ldots, d_N\}$ of which each d_n $(1 \leq n \leq N)$ is independent and identically distributed, the log probability of D under this model is given by

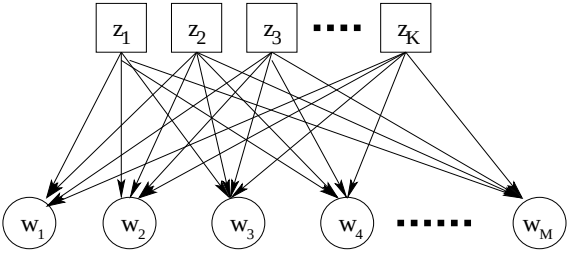

Fig. 1. The multiple-cause networks

$$\log P(D|\theta) = \sum_{n=1}^{N} \log P(d_n|\theta)$$
$$= \sum_{n=1}^{N} \log \left[\sum_{\mathbf{z}} P(\mathbf{z}|\theta) P(d_n|\mathbf{z}, \theta) \right], \quad (1)$$

where θ is the parameter set of the model, and $\mathbf{z}$ is one of all the possible combinations of topics, for example $\mathbf{z} = (1, 0, 1, \ldots, 0)$. Considering the network structure and dependency separation [5] in Figure 1, it is assumed that individual words in a document are conditionally independent given a specific topic configuration. That is,

$$P(d_n|\mathbf{z}, \theta) = \prod_{m=1}^{M} P(w_m|\mathbf{z}, \theta). \quad (2)$$

If we encode documents using binary word vectors considering only the presence or absence of words in documents, the variable w_m is a binary variable. There are several alternatives for calculating the probability $P(w_m|\mathbf{z})$, including sigmoid function [1][5], noisy-OR function [10], and competitive function [2]. For the present work, we adopt the competitive function for the activation function. This function is given by

$$P(w_m = 1|\mathbf{z}) = 1 - \frac{1}{1 + \sum_k \theta_{km} z_k}, \quad (3)$$

where θ_{km} is non-negative and interpreted as the contribution to the odds that if z_k is 1 then w_m is 1. And $P(w_m = 0|\mathbf{z})$ is $1 - P(w_m = 1|\mathbf{z})$. Details about the derivation of this function are referred to [2].

In the above settings, however, each document can be generated in exponentially many ways with the number of possible latent topics. The computational costs considering all of these explanations can make standard maximum likelihood approaches such as EM algorithm intractable [1][5]. With 20 possible topics, for example, we must consider 2^{20} $(> 1,000,000)$ configurations for each document. Therefore generally some approximations have been used, and we choose the *Helmholtz machine* [1] for this.

2.2 The Helmholtz Machines

The Helmholtz machine is a connectionist system with multiple layers of neuron-like stochastic processing units [1], and it can be used to ease the process of learning and inference in multiple-cause networks. The log-likelihood given in the Equation (1) can be lower-bounded by

$$\log P(D|\theta) = \sum_{n=1}^{N} \log \left[\sum_{\mathbf{z}} Q(\mathbf{z}|d_n) \frac{P(\mathbf{z}|\theta) P(d_n|\mathbf{z}, \theta)}{Q(\mathbf{z}|d_n)} \right]$$
$$\geq \sum_{n=1}^{N} \sum_{\mathbf{z}} Q(\mathbf{z}|d_n) \log \frac{P(\mathbf{z}|\theta) P(d_n|\mathbf{z}, \theta)}{Q(\mathbf{z}|d_n)}, \quad (4)$$

where $Q(\mathbf{z}|d_n)$ is an approximation to the true posterior probability $P(\mathbf{z}|d_n, \theta)$. If this lower bound is maximized, the true log likelihood can be approximately maximized [1].

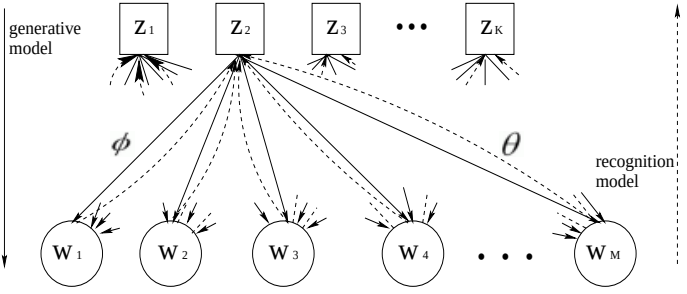

Fig. 2. A two-layer Helmholtz machine. The solid lines represent the top-down generative connections and the dotted lines represent the bottom-up recognition connections

The Helmholtz machine introduces a recognition network to estimate the distribution $Q(\mathbf{z}|d_n)$. So the machine is composed of a pair of networks. Figure 2 shows a simple Helmholtz machine with one input layer and one hidden layer.

The newly introduced recognition network is implemented by bottom-up connections ϕ, and the $Q(\mathbf{z}|d_n)$ is estimated by a parametric form $Q(\mathbf{z}|d_n, \phi)$. Since individual latent nodes in the recognition network are conditionally independent given all the values of input nodes, the probability is represented simply by

$$Q(\mathbf{z}|d_n, \phi) = \prod_{k=1}^{K} Q(z_k|d_n, \phi). \tag{5}$$

And the probability $Q(z_k|d_n, \phi)$ is calculated by

$$Q(z_k|d_n, \phi) = \frac{1}{1 + \exp\left(-\sum_{m=0}^{M} \phi_{mk} w_m\right)}. \tag{6}$$

where w_0 is 1, and ϕ_{0k} is a recognition bias to z_k. Using this approximation, we can simplify and accelerate the procedure of estimating posterior probabilities for documents.

3 Applying to Text Documents

The Helmholtz machine is basically a binary-valued Bayesian network, so it needs some modifications to handle the document of which each element is the number of occurrence of the corresponding word. In the work of [4], they have modelled the observed units as discrete Poisson random variables, by which they

have tried to automatically discover stochastic firing patterns in large ensembles of neurons. In our experiments for text documents, however, it didn't work well compared with the original binary Helmholtz machines. Rather, We utilize the idea of the *replica* in [11] where the restricted Boltzmann machines with replicas have been applied to the face recognition problem. We assume that the input nodes can have multiple *replica*s and all of them have the identical weights to and from latent nodes. The number of replicas of an input node is set to the frequency count of the corresponding word in a document.

In recognition of a document d, activities of latent topics are determined according to the probabilities given by the Equation (6). If the mth word occurs l times in the document, l binary replicas are introduced for the corresponding node. In practice this is implemented simply by setting $w_m = l$. The model is fitted using a stochastic algorithm, the *wake-sleep* algorithm [1][6]. Learning in the wake-sleep algorithm consists of interleaved iterations of two phases, the *wake* phase and the *sleep* phase. In relation to the generalized EM algorithm, the wake phase plays the role corresponding to the generalized M-step and the sleep phase corresponding to the generalized E-step [5].

Table 1 describes the learning algorithm of the Helmholtz machine when applied to text documents. This algorithm proceeds in the same way as described in [6] except for two differences. First, the activation probabilities of the latent nodes are estimated considering the frequency information of words. In this way, in estimating the latent topics, words which appear many times in a document are distinguished from those with just one occurrence. Second, activation probabilities of input nodes in the generative network are given by *competitive rule* as indicated in Section 2.1. Especially if the word count is greater than 1, this rule is applied to its binary-valued replica. So in the update rule for the generative weight θ_{km} from the kth latent node to mth input node, s_m is a binary value, that is 0 or 1.

4 Experiments

4.1 Topic Words Extraction

We have performed some experiments on topic extraction with a subset of TDT-2 collection. The subset contains topics which have relatively many relevant documents , resulting in 10,685 documents in total. Stop words in a standard stop word list have been eliminated, and words which occur in at least 10 documents have been selected. For the ease of interpretation of results, no stemming or further preprocessing has been performed. Finally, the resulting vocabulary size is 12,467.

We have trained Helmholtz machines with varying number of latent nodes. Table 2 shows 9 latent factors extracted with 32 latent factors. For each topic z, all the words are sorted according to their probabilities $P(w|z)$, and the top 15 words are shown. It can be seen that words in the same factor are semantically related or, in a weak sense, refer to the same topic.

Table 1. The learning procedure based on the wake-sleep algorithm

---

INPUT: (**D**, ϕ, θ, η, N, M, K)
D is the set of text documents. Each document $d_n = (w_1, w_2, \cdots, w_M)$ is represented by a numeric vector, where w_m is the frequency of the word in the document.
η is the learning rate.
ϕ is the set of parameters in the recognition model, and θ is the set of parameters in the generative model.
N and M are the number of documents and the vocabulary size, respectively, and K is the number of latent topics.

Iterate the two phases until the algorithm converges.

Wake phase

1. A document d is clamped on the input nodes. Using the recognition model, the latent topics are sampled from their conditional probabilities calculated by the Equation (6).
2. θ is updated to make the latent topics picked above more likely for d.

$$\theta_{km} = \theta_{km} + \eta \frac{1 - p_m}{p_m}(s_m - p_m)s_k \qquad (1 \le k \le K, \quad 1 \le m \le M)$$
$$\theta_{0k} = \theta_{0k} + \eta(s_k - p_k) \qquad \text{(biases to the latents)} \qquad (1 \le k \le K)$$

where p_k and p_m are the activation probabilities for the latent topics and words, respectively, on the generative model. s_k and s_m are the actual values of 0 or 1.

Sleep phase

1. In the generative model, the latent topics are randomly selected according to the following probabilities.

$$p_k = \frac{1}{1 + \exp(-\theta_{0k})}$$

From this, a "*fantasy*" document is generated in input layer.
2. ϕ is updated to make this fantasy case more likely.

$$\phi_{mk} = \phi_{mk} + \eta s_m (s_k - p_k) \qquad (1 \le k \le K, \quad 1 \le m \le M)$$

where p_k is the activation probability for kth latent topics on the recognition model. s_k and s_m are the actual value of 0 or 1.

Table 3 shows three latent factors extracted with 64 latent factors. With this increased factors, it is shown that the topic on "*winter Olympics*", the fourth in Table 2, is represented by three more specific latent factors. This shows the properties of multiple-cause models on text documents where several semantic factors can be combined to represent documents. In examining the results in

Table 2. Topic word sets from TDT-2 corpus. 15 most probable words are shown for each topic word set in decreasing order

15 most probable words for each latent generator
tobacco, smoking, gingrich, newt, trent, republicans, congressional, republicans, attorney, smokers, lawsuit, senate, cigarette, morris, nicotine
iraq, weapons, united, saddam, military, iraqi, inspectors, hussein, security, baghdad, nations, inspections, gulf, destruction, war
lewinsky, monica, president, starr, clinton, house, white, counsel, independent, jury, investigation, sexual, kenneth, grand, relationship
olympics, olympic, games, nagano, winter, gold, medal, men, team, skating, women, athletes, ice, ski
israeli, minister, israel, peace, prime, bank, netanyahu, palestinian, west, secretary, talks, arafat, benjamin, united, albright
india, pakistan, pakistani, delhi, hindu, vajpayee, nuclear, tests, atal, kashmir, indian, janata, bharatiya, islamabad, bihari
suharto, habibie, demonstrators, riots, indonesians, demonstrations, soeharto, resignation, jakarta, rioting, electoral, rallies, wiranto, unrest, megawati
market, stock, prices, percent, points, asia, asian, fell, index, investors, rose, stocks, financial, markets, analysts
pope, cuba, visit, paul, john, caban, castro, fidel, havana, communist, cubans, church, human, catholic, pontiff

Table 3. Topics on the winter Olympics : "ice hockey", "skating", "general & ski"

"*ICE HOCKEY*"	team, hockey, ice, canada, game, olympic, players, goal, tournament, league, scored, goalie, coach, round, victory, national, czech, period, nhl, $\cdots$, puck, stick, $\cdots$
"*SKATING*"	skating, figure, program, olympic, world, champion, skate, short, competition, lipinski, judges, medal, tara, triple, ice, kwan, jumps, skater, michelle, performance, $\cdots$
"*GENERAL & SKI*"	won, olympics, winter, games, nagano, world, race, medal, gold, silver, team, $\cdots$, ski, finish, event, final, slalom, snow, $\cdots$

detail, we have found some interesting facts. In Table 4, the first two columns represent the different contexts where the word '*race*' is used: arms *race* and *race* in sports like ski. And the last two columns represent the different usages of another word '*court*': *court* related with law and *court* related with sports like basketball. In this way, the different meanings or usages of the same word can be differentiated. Similar experimental results and interpretation for text documents with *PLSA* and *NMF* are presented in [7][8], respectively.

Table 4. Four selected topics from 64 latent factors for the subset of TDT-2 Corpus

"nuclear race"	"winter Olympics"	"legal affair"	"basketball"
india	won	case	jordan
nuclear	olympics	***COURT***	bulls
tests	winter	judge	jazz
pakistan	games	law	nba
weapons	nagano	federal	finals
hindu	world	attorney	basketball
arms	***RACE***	legal	***COURT***
RACE	medal	justice	pippen
sanction	gold	lawyers	points
security	silver	supreme	phil
delhi	team	evidence	teams
nationalist	lillehammer	trial	game

4.2 Document Clustering

We also experimented for the document clustering by the learning of Helmholtz machines. This might provide, though indirect, some quantitative view on the performance besides the qualitative results of the previous section. The dataset contains 1,069 documents of 4 topics from TREC-8 ad-hoc task data, including '*Foreign minorities, Germany*'(ID 401), '*Estonia, economy*' (ID 434), '*inventions, scientific discoveries*' (ID 439), and '*King Husayn, peace*' (ID 450). Stop words have been removed and the words with at least 5 document frequency have been selected, which result in 8,828 words. The number of latent nodes were set to four and the algorithm was run ten times with random initializations.

The clustering is performed by two methods of estimating posterior probabilities, one is by $Q(z|d)$ and the other by $P(z|d)$. Given a document d, $Q(z|d)$ is estimated simply using the recognition network , and $P(z|d)$ is by the conjugate

Table 5. Confusion matrix for a subset of *TREC* ad-hoc task data. Each document d is assigned to the most active cluster according to $Q(z|d)$ and $P(z|d)$. The bottommost row shows the average error rates across the runs where the 4 topics were relatively well separated, and the numbers of such "successful" runs are shown in the parentheses

	HM with frequency								HM with binary							
	$Q(z\|d)$				$P(z\|d)$				$Q(z\|d)$				$P(z\|d)$			
Topic ID	401	434	439	450	401	434	439	450	401	434	439	450	401	434	439	450
401	**297**	2	0	1	**297**	3	0	0	**284**	12	2	2	**285**	12	3	0
434	2	**343**	2	0	0	**346**	1	0	0	**338**	7	2	0	**333**	13	1
439	1	4	**124**	0	0	1	**128**	0	8	19	**101**	1	0	31	**98**	0
450	2	1	0	**290**	0	4	1	**288**	0	3	5	**285**	1	2	2	**288**
Error rates	$1.85 \pm 0.45\%$ (8)				$1.68 \pm 0.49\%$ (8)				$6.92 \pm 0.97\%$ (5)				$6.95 \pm 0.85\%$ (5)			

gradient method using the generative network as in [9][10]. The predicted topic $\hat{t}$ of the document d is that with the highest $Q(z_k|d)$ or $P(z_k|d)$, that is,

$$\hat{t} = \text{argmax}_k Q(z_k|d) \quad \text{or} \quad \text{argmax}_k P(z_k|d). \tag{7}$$

Table 5 shows the confusion matrix for the case with the best clustering result and the average performances across the runs where the topics are relatively well separated. As can be seen in the results, there are no significant differences between $Q(z|d)$ and $P(z|d)$ in the clustering performance. In terms of inference for a given document, however, the estimation of $Q(z|d)$ required much less computational costs than $P(z|d)$ in CPU time, *0.02 sec/doc* and *0.98 sec/doc* on a Linux system with AMD Athlon 1 GHz CPU, respectively. So, the approximation of multiple-cause networks by the architecture of Helmholtz machines could provide fast inference for newly presented documents, especially for those with high dimensionality.

The right part in Table 5 shows the clustering result when documents are represented by the binary encoding which considers only whether a word occurs in a document or not. In this case the performance is poor compared to the result using word frequency information, especially for ID 439 about scientific discoveries. From this, we argue that, though it is not utilized directly in the parameter estimation of the model but only in estimating active latent factors, the word frequency information helps produce better topic decomposition.

5 Conclusions

In this paper, we have presented a multiple-cause network based approach for topic decomposition of text documents. To ease the learning and inference in the network, we utilized the approximation given by Helmholtz machines. The competitive function has been used as the activation function for input nodes in the generative network, and the word frequency information was incorporated into the recognition network when estimating likely latent topics for each document.

In the experiments on TDT-2 collection, we have presented the topic decomposition results with varying number of latent topics and have shown the characteristics of the multiple-cause model for text documents. In the document clustering experiment, we estimated likely latent topics both on the recognition network and on the generative network. The former has provided much faster inference than the latter, without significant deterioration of the performance. And though confined to estimating latent topics for given documents, the incorporation of the word frequency has been experimentally shown to be helpful for the analysis of text documents.

Acknowledgments. This work was supported by the Korean Ministry of Science and Technology under the BrainTech Project and by the Korean Ministry of Education under the BK21-IT Program. Yuseop Kim is supported by Brain Korea 21 project performed by Ewha Institute of Science and Technology.

References

1. Dayan, P., Hinton, G.E., Neal, R. M., Zemel, R. S.: The Helmholtz machine. Neural Computation **7** (1995) 889–904
2. Dayan, P., Zemel, R.S.: Competition and multiple cause models. Neural Computation **7** (1995) 565-579
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science. **41** (1990) 391–407
4. deSa, V.R., deCharms, R.C., Merzenich, M.M.: Using Helmholtz machines to analyze multi-channel neuronal recordings. Advances in Neural Information Processing Systems 10 (1998) 131–137
5. Frey, B.J.: Graphical Models for Machine Learning and Digital Communication. The MIT Press (1998)
6. Hinton, G.E., Dayan, P., Frey, B.J., Neal, R.M.: The wake-sleep algorithm for unsupervised neural networks. Science **268** (1995) 1158–1161.
7. Hofmann, T.: Probabilistic latent semantic indexing. Proceedings of the 22th International Conference on Research and Development in Information Retrieval (SIGIR) (1999) 50–57
8. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788-791
9. Sahami, M., Hearst, M., Saund, E.: Applying the multiple cause mixture model to Text Categorization. Proceedings of the 13th International Conference on Machine Learning (1996) 435–443
10. Saund, E.: A multiple cause mixture model for unsupervised learning. Neural Computation **7** (1995) 51–71
11. Teh, Y.W., Hinton, G.E.: Rate-coded restricted Boltzmann machines for face recognition. Advances in Neural Information Processing Systems 13 (2001) 908–914