Using Stochastic Helmholtz Machine for Text Learning

Jeong-Ho Chang and Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University
San 56-1 Shillim-dong Gwanak-gu, Seoul 151-742, Republic of KOREA
Email: jhchang@scai.snu.ac.kr, btzhang@scai.snu.ac.kr

Abstract

We present an approach for text analysis, especially for topic words extraction and document classification, based on a probabilistic generative model. Generative models are useful since they can extract the underlying causal structure of data objects. For this model, a stochastic Helmholtz machine is used and it is fitted using the wake-sleep algorithm, a simple stochastic learning algorithm. Given a document set, the Helmholtz machine tries to capture the correlation of the words used in the set, thus can extract various semantic features for a set of documents. We present some experimental results on topic words extraction for TDT-2 and TREC-8 ad-hoc data sets. And for another real-world document set, 20 Newsgroup collection, a categorization is performed and the performance is compared with that of naïve Bayes classifier, another simple generative model. Additionally, we present a preliminary work to make Helmholtz machines more appropriate for processing text documents.

Keywords: Helmholtz machine; multiple-cause model; topic word; text categorization; topical clustering.

1 Introduction

With the popularity of search on the Internet and the explosion of the information in text, the demand for automatic text analysis (including indexing, clustering, categorization, and so on) is also increasing. Recently many machine learning techniques have been proposed for such purposes. In this paper, we present an approach for text analysis based on density estimation of text documents by probabilistic generative models.

Generative models are popular choices to extract the structure in a data presented as a vector [4] and can be used to estimate the density of a particular observation. When a generative model has latent variables that are not directly observed, these variables can be thought of as underlying causes which responsible for generating a data. In many cases, it makes sense to postulate that data naturally arise from the consequences of cooperative activities of a few of these possible generators.

Our work is based on the above point of view on text documents. For text documents, data are thought of as documents and hidden causes as the underlying topic word sets for the documents. In other words, the underlying latent word sets can be considered as keyword sets, and have an important role of generating documents. In this way, we can think of a document as an output of appropriate generative process. Given a document we can estimate the probability that the document would be produced through the generative model.

Generally these generative processes are modeled using graphical representation, and graphical models are useful tools for representing generative processes in probabilistic manner. If the structure of a model become complicated, however, the learning and inference in a graphical model are likely to be intractable. Fortunately, during the recent years some approximation methods have been proposed to make the learning and inference more tractable. The Helmholtz machine [2] is one of such methods. It is simple to implement and has shown a relatively good results in such problems as bar image analysis [5][6] and handwritten digit classification [5]. In this paper, we utilize Helmholtz machines for analyzing text documents.

2 The Helmholtz Machine

The Helmholtz machine is a connectionist system with multiple layers of neuron-like stochastic processing units connected hierarchically [1][2]. It has been used to ease the process of learning and inference in a binary-valued multiple-cause networks and hierarchical networks. Assume that a set of data **D** is given, and it's underlying distribution is estimated by a multiple-cause network with a set of causes **z** and a parameter set $\mathbf{\Theta}$. If each sample of **D** is independently and identically distributed, the log probability of **D** under this model is given by

$$\begin{aligned}\log P(\mathbf{D}\,|\,\mathbf{\Theta}) &= \sum_{i=1}^{N} \log P(d_i\,|\,\mathbf{\Theta}) \\ &= \sum_{i=1}^{N} \log\left(\sum_{\mathbf{z}} P(d_i, \mathbf{z}\,|\,\mathbf{\Theta})\right) \end{aligned} \qquad (1)$$

where **z** represents possible underlying causes for a pattern d_i. Since each data can be generated in exponentially many ways, the computational costs considering all of these can make standard maximum likelihood approaches such as

EM algorithm intractable. Therefore, it is usual that some approximations are introduced.

Helmholtz machines introduce, in addition to a generative network, a recognition network to compute an approximate distribution over the hidden causes **z** for each data pattern. Figure 1 shows a simple Helmholtz machine with one input layer and one hidden layer. The generative model is implemented by top-down connections **Θ** and the recognition model is implemented by bottom-up connections **Φ**. The nodes are stochastic binary nodes, and those in the same layer are conditionally independent one another, given the values of the nodes leading into those.

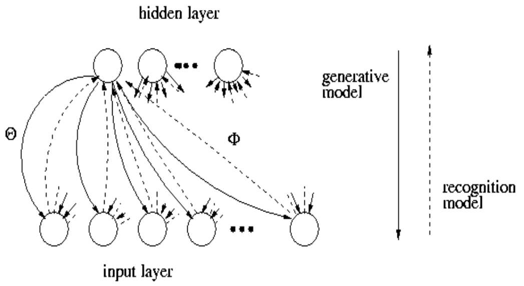

Figure 1. A Helmholtz machine

In the recognition network, the probability that the *j*th unit in layer *l* has activity s_j=1 is the function of the activities s_i of the units in layer (*l*-1), and this is given by

$$q_j(\mathbf{\Phi}, \mathbf{s}^{l-1}) = \sigma\left(\sum_{s_i \in (l-1)} s_i \phi_{ij} \right) \tag{2}$$

In the generative network, the probability that the *j*th unit in layer *l* has activity s_j=1 is the function of the activities s_i of the units in layer (*l*+1), and given by

$$p_j(\mathbf{\Theta}, \mathbf{s}^{l+1}) = \sigma\left(\sum_{s_i \in (l+1)} s_i \theta_{ij} \right) \tag{3}$$

where $\sigma(x) = 1/[1+\exp(-x)]$ [2].

By iteratively adjusting all the weights in both generative network and recognition network, we can maximize a lower bound on the log likelihood for a give data set. A simple stochastic algorithm, called the *wake-sleep algorithm* [6], is used for the adjustment.

3 Applying the Helmholtz machines to Text Documents.

We intend to capture the underlying regularities or causal structures in a text corpus by finding coherent sets of words within a high-dimensional document space. A document is encoded as *n* dimensional binary vector, where *n* is the vocabulary size for a document set. Thus each entry of the vector represents whether the corresponding word exists in the document.

In applying to text documents, we use two-layer Helmholtz machines as depicted in Figure 1. Both hidden nodes and input nodes are binary as proposed originally in [2]. For input nodes on the generative network, however, we use *competitive* activation functions proposed in [3] in stead of sigmoid activation function, and these are used to estimate the probability of being "*on*" of the input nodes. This function is given by

$$P_j(\Theta, \mathbf{s}^1) = 1 - \frac{1}{1 + \sum_i s_i^1 \theta_{ij}} \tag{4}$$

where each weight on the corresponding connection from the latent node to input node is non-negative and interpreted as the contribution to the odds that if s_i is 1 then s_j is 1. These non-negativity properties permit the combination of the latent topics to represent a document. But, only additive combinations of the latent topics are allowed and no subtraction can occur. This leads to the easier interpretation of the semantic analysis results. For the remaining nodes, sigmoid activation function is used in Equation 3.

In the next two sections, we present more specific settings of Helmholtz machines for various experiments. And in Section 6, a slight modification that permits Helmholtz machines to utilize word count information is introduced.

4 Topic Words Extraction

4.1 Data Sets and Experimental Setup

Two real-world document corpora are experimented with, TREC-8 ad-hoc data set and TDT-2 collection. Among TREC-8 data set, we have selected four topic sets that have relatively many relevant documents, 1,069 documents in total. These topics include: *foreign minorities in Germany* (*ID 401*), *economy in Estonia* (*ID 434*), *inventions and scientific discoveries* (*ID 439*), and *King Husayn and peace in the Middle East* (*ID 450*). In the case of TDT-2 collection, a subset of the data set has been selected of which each topic has more than 100 documents, 6,169 documents in total. Since the topics and events of these document collections will be familiar for most readers, it will be easier to verify the resulting sets of topic words.

All uppercase letters in a document have been converted to lowercase, so effectively capitalization is ignored. Stop words in a standard stop word list have been eliminated, and also words that occur no more than 5 times have been removed. For the ease of analysis of the result, no stemming or further preprocessing has been performed. The resulting vocabulary sizes were 8,828 for the TREC-8 ad-hoc corpus and 13,213 for the TDT-2 corpus. Finally each document has been encoded using the binary features, 1 for the existence of a word *w* and 0 for the absence of the word.

For the TREC-8 collection, as an initial experiment to test the viability of the Helmholtz machine to text documents, the number of the latent factors has been set to four which is the same as the number of topics in the data set. By this setting, each document is assumed to be represented by only one latent factor. For TDT-2 collection, we have experimented with various number of latent factors, 16, 32, 64 respectively.

4.2 Experimental Results

Table 1 shows the four factors extracted by the Helmholtz machine for TREC-8 data set. In each latent factor, it can be seen that the related words of each topic have been grouped together. Though the Helmholtz machine is identified as a kind of multiple-cause model where each data is represented by more than one factor, we can see that it works well when the documents well separated in their topics. Experiments on the document clustering based on this result are presented Section 6.

Table 1. Four factors from TREC-8 ad-hoc collection. The words w of each factor z are ordered according to P(w|z).

401	germany, german, asylum, foreigners, wing, minister, social, foreign, union, Turkish
434	estonia, economic, foreign, trade, Estonian, government, country, state, Russia, Tallinn
439	company, technology, development, market, production, research, make, patent, work, cost
450	king, jordan, peace, israel, jordanian, israeli, talks, arab, minister, east

Table 2 shows some of the latent factors for TDT-2 collection, extracted with 16 latent factors. As in experiment for TREC-8 data set, we can see that the related words on the same topic have been grouped together.

Table 2. Some topic word sets from TDT-2 corpus.

warplane, airline, saudi, gulf, wright, soldiers, yitzhak, tanks, stealth, kurds
olympics, nagano, olympic, winter, medal, hockey, athletes, cup, slalom, medals
netanyahu, palestinian, arafat, israeli, yasser, kofi, annan, benjamin, mideast, gaza, Jerusalem
imf, monetary, currencies, currency, rupiah, singapore, bailout, traders, markets, Thailand
pope, cuba, cuban, embargo, castro, lifting, cubans, havana, alan, invasion

Table 3 show three latent factors accomplished with 64 latent factors. With this increased factors, it is shown that the topic on "*winter Olympics*", the item on the second row in Table 2, is decomposed into 3 subtopics. This shows the properties of the Helmholtz machine as a multiple-cause model on text documents where each document is represented by combining several of the semantic features.

Table 3. Topics on the winter Olympics: "Skating", "Ice hockey", "General & Ski".

SKATING	skating, figure, program, olympic, champion, skate, short, competition, lipinski, judges, medal, tara, triple, ice, kwan, jumps, …
ICE HOCKEY	team, hockey, ice, canada, game, olympic, players, goal, tournament, leaguer, scored, …, puck, stick, …
GENERAL & ETC	won, olympics, winter, games, nagano, world, race, medal, gold, silver, …, ski, finish, event, final, slalom, …

Additionally, in examining the result in more detail, we have found an interesting fact. This is shown in Table 4. For each factor *z*, a word *w* is in decreasing order of the conditional probability *P*(*w*|*z*). The first two columns represent the different contexts where the word 'bank' is used: *bank* related with finance and *bank* related with river or shore. And the last two contexts represent the different usages of another word 'race': arms *race* and *race* in games like ski. In this way, the different meanings or usages of the same word can be differentiated by Helmholtz machines. Similar experimental results with PLSA and NMF are presented in [7][9].

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

"financial crisis"	"Israel"	"nuclear race"	"winter Olympics"
asia	israel	india	Won
economy	palestinian	nuclear	olympics
percent	peace	pakistan	winter
financial	netanyahu	tests	games
market	process	arms	nagano
crisis	arafat	***RACE***	world
currency	***BANK***	test	***RACE***
dollar	benjamin	weapons	medal
japan	yasser	indian	gold
…	…	…	silver
BANK	talks	security	…
….	…	…	athletes

5 Text Categorization

This section presents the results of text document categorization by Helmholtz machines.

5.1 Data Set and Experimental Setup

In our text categorization experiment, 20 UseNet newsgroup data set is used. This data set, collected by Ken Lang [8], contains 20,000 articles evenly divided among 20 discussion groups [8][11].

As in previous experiments in Section 4, a document is represented by a vector of binary attributes indicating which words occur and do not occur in the document. In addition to the preprocessing described in Section 4, a further step has been introduced. We have selected 6,000 informative words by pruning words after calculating the information gain of each word. The information gain for a word is calculated as follows.

$$G(w) = -\sum_{c} P(c) \log P(c) + P(w) \sum_{c} P(c \mid w) \log P(c \mid w) + P(w) \sum_{c} P(c \mid \neg w) \log P(c \mid \neg w) \quad (5)$$

Where $G(w)$ is the information gain of a word w, $P(c)$ is the probability of a category, $P(w)$ is the probability of a word, and $P(c \mid w), P(c \mid \neg w)$ are the probabilities of a category c when there exists w and does not exist, respectively.

The distribution of each of the 20 classes of newsgroup data has been modeled using a separate stochastic Helmholtz machine with 6,000 binary input nodes and 1 hidden layer of 10 binary nodes. As in [5], once the 20 Helmholtz machine have been fitted, classification of a test document, *d*, proceeds by estimating the class likelihood *P*(*d*|*c*), for each machine. This likelihood is estimated using 10 recognition sweeps. And Bayes' rule for the posterior probability of class *c* given a document *d* is calculated to produce a soft classification decision

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{\sum_{c'} P(d \mid c')P(c')}, \quad (6)$$

where P(*c*) is the prior probability of each category *c* and *P*(*d*|*c*) is the conditional probability that a document *d* is generated given a category *c*. In this experiment, P(*c*) is assumed to be equal to the empirical probability, $\widetilde{P}(c)$ (=1/20). Thus,

$$P(c \mid d) = \frac{P(d \mid c)}{\sum_{c'} P(d \mid c')}, \qquad c \in \{0, 1, \ldots, 19\} \quad (7)$$

Finally, a hard decision c* is made by choosing the best class which satisfies

$$c^* = \arg\max_c P(c \mid d) \quad (8)$$

5.2 Experimental Results

The results of categorization are shown in Table 5. For performance comparison, experimental results by naïve Bayes classifier are also presented. The naïve Bayes classifier assumes that all words of a document are independent of one another given the category of the document. There are two different generative models with this "naïve Bayes assumption", that is multi-variate Bernoulli event model and multinomial event model [10]. The multi-variate Bernoulli event model specifies a document in a binary vector over the space of words, like our Helmholtz machine model. The multinomial event model specifies that a document be represented by the set of word occurrences of each word in the document. In this event model, the number of occurrences of each word in the document is captured.

Table 5. Results of text categorization on 20 electronic newsgroup data.

	naïve Bayes classifier		Helmholtz machine
	multi-variate Bernoulli	multinomial	
Accuracy	73.95 %	76.89 %	81.26 %

As indicated in [10], on 20 UseNet data set with high vocabulary size, the performance of the multi-variate Bernoulli event model is worse than that of the multinomial event model. By introducing the correlation between words in a document, however, Helmholtz machine achieved slightly better results compared to the multinomial event model in spite that it uses binary document vector. This shows that it can be helpful in text categorization to capture the correlation among words in a document.

6 Using Word Count Information in Helmholtz Machine : A Preliminary Work

In the experiments in Section 4 and Section 5, only binary information about the word existence in a document has been used. For such problems as text document clustering, however, it can be helpful to consider the number of words occurring in documents [12]. In this section, we present a preliminary work for using the word count information in the Helmholtz machine.

6.1 A Modification of Helmholtz Machine

In the work of [4], they have proposed to model 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 experiment for text document, however, it didn't work well compared with the original binary Helmholtz machines. Rather, we assume that each observed unit could have multiple replicas, all of which have identical weights to and from the hidden units. This approach is similar to that of [14], where binary-valued, restricted Boltzmann machine with replicas has been applied to the problem of face recognition. But unlike their work, we don't introduce replicas for hidden nodes. In our work, the number of replicas of a visible node is the same as the number of occurrences of the corresponding word. With this replica trick, we can naturally utilize the word frequency information without changing the inference and learning procedures.

In recognition for a document *d*, the activities of the hidden causes are determined according to the probability given by the Equation 3. If a word corresponding to the *i*th node in the input layer occurs *m* times in the document, *n* replicas for the node are introduced. In other words, the node split into m binary-valued nodes of which all the activities are 1. Figure 2 shows a two-layer Helmholtz machine with replicas. In practice, this is implemented simply by setting $s_i=m$ in Equation 3. The generative network proceeds in the same way as the binary Helmholtz machine.

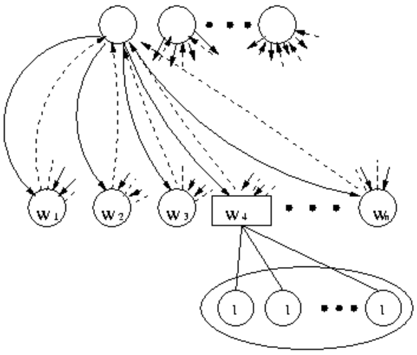

Figure 2. The structure of a Helmholtz machine with replicas

To test the properness of the modified Helmholtz machine for text documents, we have experimented on topical clustering of documents and compared the result with that of the machine which considers binary information only.

6.2 Topical Clustering

We have experimented with two data collection used in the previous experiments, TREC-8 ad-hoc data set and Newsgroup collection. For the latter, we have selected two disjoint subsets, which are about *recreation* and *science* respectively. The details are described in Table 6.

Table 6. Subsets of Newsgroup corpus. Documents in each set have been randomly chosen from the corpus.

Data Set	Topics	# of doc in each group	# of doc in total
Recreation	Autos motorcycles baseball hockey	500	2,000
Science	Crypt electronics medical space	500	2,000

Each data set has been preprocessed as in Section 4. For the two Newsgroup data, after additional feature selection mechanism used in [13], we selected 2,000 words with the highest contributions to the mutual information between the words and the documents. Finally each document is represented employing "bag-of-words" or numeric vector space representation [12].

Table 7 shows overall clustering performance on the three data sets.

Table 7. Clustering errors of Helmholtz machines on three data set.

	HM_*numeric*	HM_*binary*
TREC ad-hoc	2.2 % (4)	7.7 % (4)
Science	14.9 % (4) 13.6 % (6)	16.2 % (6)
Recreation	9.7 % (5)	11.6 % (6)

After the model is fitted, each document d has been assigned to the cluster c^* with the highest posterior probability $P(z_k|d)$ among all latent factors. That is,

$$c^* = \arg\max_{z_k} P(z_k \mid d) \tag{9}$$

Each probability has been approximated using the recognition network, and HM_*numeric* and HM_*binary* represent the Helmholtz machines with numeric vector representation and binary vector representation respectively. Note that HM_*numeric* always outperforms HM_*binary*.

For "recreation" data set, the clustering result was not satisfactory with just four latent nodes, and there have been consistent confusions between topics 'baseball' and 'hockey', or between 'autos' and 'motorcycles". When added another latent node, the experiment has shown some reasonable performance. The numbers in parentheses in the Table 6 represent the number of latent nodes of the corresponding Helmholtz machine. Among these five latent nodes, one node has no topic taking the significant majority, and a close examination of the most probable words for the latent node shows that it doesn't represent well any particular topic. It just contains the common words across all the electronic news articles. The details are shown in Table 8. When clustering documents, we simply ignore this residual node.

Table 8. The 10 most probable words of each latent factor for "recreation" data set. Note that the last factor represents no specific topic.

0	bike, ride, good, riding, motorcycle, bmw, bikes, ama, road, rider
1	game, time, year, baseball, play, good, games, league, season, team
2	team, hockey, season, year, nhl, game, pittsburgh, toronto, play, fun
3	car, engine, good, cars, drive, people, speed, ford, make, price
4	world, time, people, good, mail, make, canada, real, read, post

For "science" data set, when it is given only four or five latent nodes, *HM_binary* has not been able to extract appropriate topic clusters, and has shown very poor performance, only 48.5 % accuracy for four and 58.6 % accuracy for five. The result of *HM_binary* for "science" data set in the Table 7 is that of being given six latent nodes. The two residual nodes have been ignored when clustering as for "recreation". Six latent factors for this data set using *HM_binary* are shown in Table 9. From these results, we can see the property of the Helmholtz machine as a multiple-cause model.

Table 9. The 10 most probable words of six latent nodes of HM_binary for "science" data set. The first two factors represent no specific topic, rather include a common words across the news articles about science.

1	write, articles, apr, university, people, time, good, phone, cs, read
2	information, program, system, time, research, list, general, years, contact, institute
3	space, nasa, gov, launch, shuttle, high, orbit, earth, moon, people
4	good, years, case, problem, medical, food, problems, doctor, find, experience
5	key, encryption, clipper, chip, government, system, keys, public, law, information
6	power, good, low, work, run, radio, line, signal, data, high

6 Conclusions

We have introduced a simple probabilistic graphical model for text analysis. The Helmholtz machine, as a generative model, can extract the underlying structure in a pattern data set, and when applied to text documents, it is able to find latent factors for a document or capture the correlation among particular words. The Helmholtz machine is simple to implement and there exists a simple stochastic learning algorithm, that is, wake-sleep algorithm. In experiments on some real-world data sets, we could see that the Helmholtz machine can extract some topic words in the documents relatively well. When applied to the problem of text categorization, the Helmholtz machine has shown some slightly better results compared to naïve Bayes classifier in spite that it is restricted to use only the binary information about the existence of words in a document.

We have presented a simple approach of utilizing word count information in a document in the Helmholtz machine. By doing so, it has been shown that the performance of the Helmholtz machine on document clustering can be improved. A further study on this method will proceed, from which we expect that we could get some better results on text analysis, including topic words extraction and text categorization.

Finally, the Helmholtz machine, like most of the graphical models, requires much computational costs if the dimension of the input space is high. Thus, considering more practical applications, more research should be done to lessen these costs. For example, more sophisticated dimensionality reduction will be a candidate for this.

Acknowledgement

This work was supported in part by the Korean Ministry of Science and Technology (BR-2-1-G-06) and by the Korean Ministry of Education under the BK21-IT Program.

References

1 Dayan, P. and Hinton, G. E. Varieties of Helmholtz machine. *Neural Networks*, 1996, 9, pp. 1385 - 1403.

2 Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. The Helmholtz machine. *Neural Computation*, 1995, 7, pp. 889 - 904.

3 Dayan, P. and Zemel, R. S. Competition and multiple cause models. *Neural Computation*, 1995, 7, pp. 565 - 579.

4 de Sa, V. R., deCharms, R. C., and Merzenich, M. M. Using Helmholtz machines to analyze multi-channel neuronal recordings. In M. I. Jordan and M. J. Kearns & S. A. Solla, editors, Advances in Neural Information Processing Systems 10, 1998, pp. 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., and Neal, R. M. The wake-sleep algorithm for unsupervised neural networks. *Science*, 1995, 268, pp. 1158 - 1161.

7 Hoffman, T. Probabilistic latent semantic indexing. In *Proceedings of the 22th International Conference on Research and Development in Information Retrieval* (*SIGIR*), 1999, pp. 50 – 57.

8 Lang, K. Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, 1995, pp. 331 - 339.

9 Lee, D. and Seung, S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, 401, pp. 788 - 791.

10 McCallum, A. and Nigam, K. A comparison of event models for naïve Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

11 Mitchell, T. M. Machine Learning. The McGraw-Hill, 1997.

12 Sahami, M. Using Machine Learning to Improve Information Access. Ph.D. thesis, Stanford University, CS Dept., 1998.

13 Slonim, N. and Tishby, N. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23th International Conference on Research and Development in Information Retrieval* (*SIGIR*), 2000, pp. 208 – 215.

14 The, Y. W. and Hinton, G. E. Rate-coded restricted Boltzmann machines for face recognition, *Neural Information Processing Systems* (NIPS), 2000