HTM: A Topic Model for Hypertexts

Congkai Sun*
Department of Computer Science
Shanghai Jiaotong University
Shanghai, P. R. China
martinsck@hotmail.com

Bin Gao
Microsoft Research Asia
No.49 Zhichun Road
Beijing, P. R. China
bingao@microsoft.com

Zhenfu Cao
Department of Computer Science
Shanghai Jiaotong University
Shanghai, P. R. China
zfcao@cs.sjtu.edu.cn

Hang Li
Microsoft Research Asia
No.49 Zhichun Road
Beijing, P. R. China
hangli@microsoft.com

Abstract

Previously topic models such as PLSI (Probabilistic Latent Semantic Indexing) and LDA (Latent Dirichlet Allocation) were developed for modeling the contents of *plain texts*. Recently, topic models for processing *hypertexts* such as web pages were also proposed. The proposed hypertext models are generative models giving rise to both *words* and *hyperlinks*. This paper points out that to better represent the contents of hypertexts it is more essential to assume that the hyperlinks are fixed and to define the topic model as that of generating words only. The paper then proposes a new topic model for hypertext processing, referred to as Hypertext Topic Model (HTM). HTM defines the distribution of words in a document (i.e., the content of the document) as a mixture over latent topics in the document *itself* and latent topics in the documents which *the document cites*. The topics are further characterized as distributions of words, as in the conventional topic models. This paper further proposes a method for learning the HTM model. Experimental results show that HTM outperforms the baselines on topic discovery and document classification in three datasets.

1 Introduction

Topic models are probabilistic and generative models representing contents of documents. Examples of topic models include PLSI (Hofmann, 1999) and LDA (Blei et al., 2003). The key idea in topic modeling is to represent topics as distributions of words and define the distribution of words in document (i.e., the content of document) as a mixture over hidden topics. Topic modeling technologies have been applied to natural language processing, text mining, and information retrieval, and their effectiveness have been verified.

In this paper, we study the problem of topic modeling for *hypertexts*. There is no doubt that this is an important research issue, given the fact that more and more documents are available as hypertexts currently (such as web pages). Traditional work mainly focused on development of topic models for plain texts. It is only recently several topic models for processing hypertexts were proposed, including Link-LDA and Link-PLSA-LDA (Cohn and Hofmann, 2001; Erosheva et al., 2004; Nallapati and Cohen, 2008).

We point out that existing models for hypertexts may not be suitable for *characterizing contents of hypertext documents*. This is because all the models are assumed to generate both words and hyperlinks (outlinks) of documents. The generation of the latter type of data, however, may not be necessary for the tasks related to contents of documents.

In this paper, we propose a new topic model for hypertexts called HTM (Hypertext Topic Model), within the Bayesian learning approach (it is similar to LDA in that sense). In HTM, the hyperlinks of hypertext documents are supposed to be given. Each document is associated with one topic distribution. The word distribution of a document is defined as a mixture of latent topics of the document *itself* and latent topics of documents *which the document cites*. The topics are further defined as distributions

* This work was conducted when the first author visited Microsoft Research Asia as an intern.

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 514–522,
Honolulu, October 2008.

of words. That means the content (topic distributions for words) of a hypertext document is not only determined by the topics of itself but also the topics of documents it cites. It is easy to see that HTM contains LDA as a special case. Although the idea of HTM is simple and straightforward, it appears that this is the first work which studies the model.

We further provide methods for learning and inference of HTM. Our experimental results on three web datasets show that HTM outperforms the baseline models of LDA, Link-LDA, and Link-PLSA-LDA, in the tasks of topic discovery and document classification.

The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 describes the proposed HTM model and its learning and inference methods. Experimental results are presented in Section 4. Conclusions are made in the last section.

2 Related Work

There has been much work on topic modeling. Many models have been proposed including PLSI (Hofmann, 1999), LDA (Blei et al., 2003), and their extensions (Griffiths et al., 2005; Blei and Lafferty, 2006; Chemudugunta et al., 2007). Inference and learning methods have been developed, such as variational inference (Jordan et al., 1999; Wainwright and Jordan, 2003), expectation propagation (Minka and Lafferty, 2002), and Gibbs sampling (Griffiths and Steyvers, 2004). Topic models have been utilized in topic discovery (Blei et al., 2003), document retrieval (Xing Wei and Bruce Croft, 2006), document classification (Blei et al., 2003), citation analysis (Dietz et al., 2007), social network analysis (Mei et al., 2008), and so on. Most of the existing models are for processing plain texts. There are also models for processing hypertexts, for example, (Cohn and Hofmann, 2001; Nallapati and Cohen, 2008; Gruber et al., 2008; Dietz et al., 2007), which are most relevant to our work.

Cohn and Hofmann (2001) introduced a topic model for hypertexts within the framework of PLSI. The model, which is a combination of PLSI and PHITS (Cohn and Chang, 2000), gives rise to both the words and hyperlinks (outlinks) of the document in the generative process. The model is useful when the goal is to understand the distribution of links as well as the distribution of words. Erosheva et al (2004) modified the model by replacing PLSI with LDA. We refer to the modified mode as Link-LDA and take it as a baseline in this paper. Note that the above two models do not directly associate the topics of the citing document with the topics of the cited documents.

Nallapati and Cohn (2008) proposed an extension of Link-LDA called Link-PLSA-LDA, which is another baseline in this paper. Assuming that the citing and cited documents share similar topics, they explicitly model the information flow from the citing documents to the cited documents. In Link-PLSA-LDA, the link graph is converted into a bipartite graph in which links are connected from citing documents to cited documents. If a document has both inlinks and outlinks, it will be duplicated on both sides of the bipartite graph. The generative process for the citing documents is similar to that of Link-LDA, while the cited documents have a different generative process.

Dietz et al (2007) proposed a topic model for citation analysis. Their goal is to find topical influence of publications in research communities. They convert the citation graph (created from the publications) into a bipartite graph as in Link-PLSA-LDA. The content of a citing document is assumed to be generated by a mixture over the topic distribution of the citing document and the topic distributions of the cited documents. The differences between the topic distributions of citing and cited documents are measured, and the cited documents which have the strongest influence on the citing document are identified.

Note that in most existing models described above the hyperlinks are assumed to be generated and link prediction is an important task, while in the HTM model in this paper, the hyperlinks are assumed to be given in advance, and the key task is topic identification. In the existing models for hypertexts, the content of a document (the word distribution of the document) are not decided by the other documents. In contrast, in HTM, the content of a document is determined by itself as well as its cited documents. Furthermore, HTM is a generative model which can generate the contents of all the hypertexts in a collection, given the link structure of the collection. Therefore, if the goal is to accurately learn and pre-

Table 1: Notations and explanations.

T	Number of topics
$\mathbf{D}$	Documents in corpus
D	Number of documents
α_θ, α_β	Hyperparameters for θ and β
λ	Hyperparameter to control the weight between the citing document and the cited documents
θ	Topic distributions for all documents
β	Word distribution for topic
b, c, z	Hidden variables for generating word
d	document (index)
$\mathbf{w}_d$	Word sequence in document d
N_d	Number of words in document d
L_d	Number of documents cited by document d
I_d	Set of cited documents for document d
i_{dl}	Index of l^{th} cited document of document d
ξ_d	Distribution on cited documents of document d
θ_d	Topic distribution associated with document d
b_{dn}	Decision on way of generating n^{th} word in document d
c_{dn}	Cited document that generates n^{th} word in document d
z_{dn}	Topic of n^{th} word in document d

Algorithm 1 Generative Process of HTM

for each document d **do**
  Draw $\theta_d \sim Dir(\alpha_\theta)$.
end for
for each word w_{dn} **do**
  if $L_d > 0$ **then**
    Draw $b_{dn} \sim Ber(\lambda)$
    Draw $c_{dn} \sim Uni(\xi_d)$
    if $b_{dn} = 1$ **then**
      Draw $z_{dn} \sim Multi(\theta_d)$
    else
      Draw $z_{dn} \sim Multi(\theta_{I_{dc_{dn}}})$
    end if
  else
    Draw a topic $z_{dn} \sim Multi(\theta_d)$
  end if
  Draw a word $w_{dn} \sim P(w_{dn} \mid z_{dn}, \beta)$
end for

dict contents of documents, the use of HTM seems more reasonable.

3 Hypertext Topic Model

3.1 Model

In topic modeling, a probability distribution of words is employed for a given document. Specifically, the probability distribution is defined as a mixture over latent topics, while each topic is future characterized by a distribution of words (Hofmann, 1999; Blei et al., 2003). In this paper, we introduce an extension of LDA model for hypertexts. Table 1 gives the major notations and their explanations.

The graphic representation of conventional LDA is given in Figure 1(a). The generative process of LDA has three steps. Specifically, in each document a topic distribution is sampled from a prior distribution defined as Dirichlet distribution. Next, a topic is sampled from the topic distribution of the document, which is a multinominal distribution. Finally, a word is sampled according to the word distribution of the topic, which also forms a multinormal distribution.

The graphic representation of HTM is given in Figure 1(b). The generative process of HTM is described in Algorithm 1. First, a topic distribution is sampled for each document according to Dirichlet distribution. Next, for generating a word in a document, it is decided whether to use the current document or documents which the document cites. (The weight between the citing document and cited documents is controlled by an adjustable hyperparameter λ.) It is also determined which cited document to use (if it is to use cited documents). Then, a topic is sampled from the topic distribution of the selected document. Finally, a word is sampled according to the word distribution of the topic. HTM naturally mimics the process of writing a hypertext document by humans (repeating the processes of writing native texts and anchor texts).

The formal definition of HTM is given below. Hypertext document d has N_d words $\mathbf{w_d} = w_{d1} \cdots w_{dN_d}$ and L_d cited documents $I_d = \{i_{d1}, \ldots, i_{dL_d}\}$. The topic distribution of d is θ_d and topic distributions of the cited documents are $\theta_i, i \in I_d$. Given λ, θ, and β, the conditional probability distribution of $\mathbf{w_d}$ is defined as:

$$p(\mathbf{w_d}|\lambda,\theta,\beta) = \prod_{n=1}^{N_d} \sum_{b_{dn}} p(b_{dn}|\lambda) \sum_{c_{dn}} p(c_{dn}|\xi_d)$$
$$\sum_{z_{dn}} p(z_{dn}|\theta_d)^{b_{dn}} p(z_{dn}|\theta_{i_{dc_{dn}}})^{1-b_{dn}} p(w_{dn}|z_{dn},\beta).$$

Here ξ_d, b_{dn}, c_{dn}, and z_{dn} are hidden variables. When generating a word w_{dn}, b_{dn} determines whether it is from the citing document or the cited documents. c_{dn} determines which cited document it

is when $b_{dn} = 0$. In this paper, for simplicity we assume that the cited documents are equally likely to be selected, i.e., $\xi_{di} = \frac{1}{L_d}$.

Note that θ represents the topic distributions of all the documents. For any d, its word distribution is affected by both θ_d and $\theta_i, i \in I_d$. There is a propagation of topics from the cited documents to the citing document through the use of $\theta_i, i \in I_d$.

For a hypertext document d that does not have cited documents. The conditional probability distribution degenerates to LDA:

$$p(\mathbf{w_d}|\theta_d, \beta) = \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta).$$

By taking the product of the marginal probabilities of hypertext documents, we obtain the conditional probability of the corpus $\mathbf{D}$ given the hyperparameters $\lambda, \alpha_\theta, \beta$,

$$\begin{aligned} &p(\mathbf{D}|\lambda, \alpha_\theta, \beta) = \\ &\int \prod_{d=1}^{D} p(\theta_d|\alpha_\theta) \prod_{n=1}^{N_d} \sum_{b_{dn}} p(b_{dn}|\lambda) \sum_{c_{dn}} p(c_{dn}|\xi_d) \\ &\sum_{z_{dn}} p(z_{dn}|\theta_d)^{b_{dn}} p(z_{dn}|\theta_{I_{dc_{dn}}})^{1-b_{dn}} \\ &p(w_{dn}|z_{dn}, \beta) d\theta. \end{aligned} \quad (1)$$

Note that the probability function (1) also covers the special cases in which documents do not have cited documents.

In HTM, the content of a document is decided by the topics of the document as well as the topics of the documents which the document cites. As a result contents of documents can be 'propagated' along the hyperlinks. For example, suppose web page A cites page B and page B cites page C, then the content of page A is influenced by that of page B, and the content of page B is further influenced by the content of page C. Therefore, HTM is able to more accurately represent the contents of hypertexts, and thus is more useful for text processing such as topic discovery and document classification.

3.2 Inference and Learning

An exact inference of the posterior probability of HTM may be intractable, we employ the mean field variational inference method (Wainwright and Jordan, 2003; Jordan et al., 1999) to conduct approximation. Let $I[\cdot]$ be an indicator function. We first define the following factorized variational posterior distribution q with respect to the corpus:

$$q = \prod_{d=1}^{D} q(\theta_d|\gamma_d)$$

$$\prod_{n=1}^{N_d} \left(q(x_{dn}|\rho_{dn}) (q(c_{dn}|\psi_{dn}) \right)^{I[L_d>0]} q(z_{dn}|\phi_{dn}) \ ,$$

where γ, ψ, ϕ, and ρ denote free variational parameters. Parameter γ is the posterior Dirichlet parameter corresponding to the representations of documents in the topic simplex. Parameters ψ, ϕ, and ρ correspond to the posterior distributions of their associated random variables. We then minimize the KL divergence between q and the true posterior probability of the corpus by taking derivatives of the loss function with respect to variational parameters. The solution is listed as below.

Let β_{iv} be $p(w^v_{dn} = 1|z^i = 1)$ for the word v. If $L_d > 0$, we have

E-step:

$$\begin{aligned} \gamma_{d_i} &= \alpha_{\theta_i} + \sum_{n=1}^{N_d} \rho_{dn} \phi_{dni} + \sum_{d'=1}^{D} \sum_{l=1}^{L_{d'}} I\left[i_{d'l} = d\right] \\ &\quad \sum_{n=1}^{N_{d'}} (1 - \rho_{d'n}) \psi_{d'nl} \phi_{d'ni} \ . \end{aligned}$$

$$\begin{aligned} \phi_{dni} &\propto \beta_{iv} \exp\left\{ \rho_{dn} E_q\left[\log\left(\theta_{di}\right)|\gamma_d\right] \right. \\ &+ \left. (1 - \rho_{dn}) \sum_{l=1}^{L_d} \psi_{dnl} E_q\left[\log\left(\theta_{I_{dl}i}\right)|\gamma_{I_{dl}}\right] \right\} \ . \end{aligned}$$

$$\begin{aligned} \rho_{dn} &= \left(1 + \left(\exp\left\{ \sum_{i=1}^{k} \left((\phi_{dni} E_q[\log(\theta_{di})|\gamma_d] \right. \right. \right. \right. \\ &- \left. \left. \left. \sum_{l=1}^{L_d} \psi_{dnl} \phi_{dni} E_q[\log(\theta_{I_{dl}i})|\gamma_{I_{dl}}] \right) \right. \right. \\ &+ \left. \left. \log\lambda - \log\left(1 - \lambda\right) \right\} \right)^{-1} \right)^{-1} . \end{aligned}$$

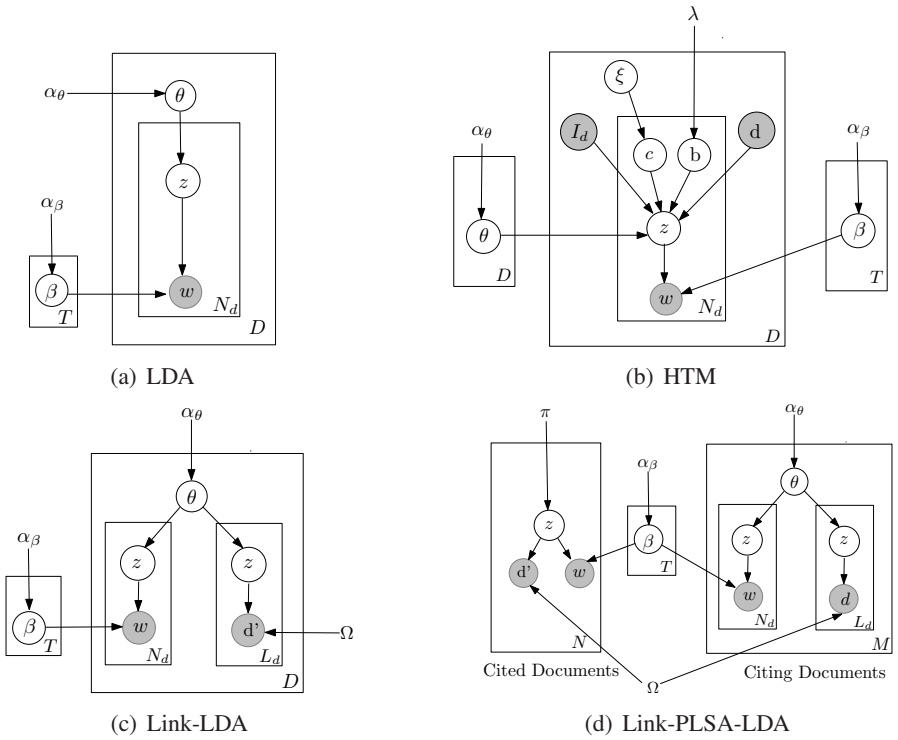

Figure 1: Graphical model representations

$$\varphi_{dnl} \propto \xi_{dl} \exp\{(1-\rho_{dn}) \sum_{i=1}^{k} \phi_{dni} E_q[\log(\theta_{I_{dl}i})|\gamma_{I_{dl}}]\}.$$

Otherwise,

$$\gamma_{d_i} = \alpha_{\theta_i} + \sum_{n=1}^{N_d} \phi_{dni} + \sum_{d'=1}^{D} \sum_{l=1}^{L_{d'}} I\left[i_{d'l} = d\right] \sum_{n=1}^{N_{d'}} (1-\rho_{d'n}) \psi_{d'nl} \phi_{d'ni} .$$

$$\phi_{dni} \propto \beta_{iv} \exp\left\{E_q\left[\log\left(\theta_{di}\right)|\gamma_d\right]\right\}.$$

From the first two equations we can see that the cited documents and the citing document jointly affect the distribution of the words in the citing document.

M-step:

$$\beta_{ij} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}.$$

In order to cope with the data sparseness problem due to large vocabulary, we employ the same technique as that in (Blei et al., 2003). To be specific, we treat β as a $K * V$ random matrix, with each row being independently drawn from a Dirichlet distribution $\beta_i \sim Dir(\alpha_\beta)$. Variational inference is modified appropriately.

4 Experimental Results

We compared the performances of HTM and three baseline models: LDA, Link-LDA, and Link-PLSA-LDA in topic discovery and document classification. Note that LDA does not consider the use of link information; we included it here for reference.

4.1 Datasets

We made use of three datasets. The documents in the datasets were processed by using the Lemur Took kit (http://www.lemurproject.org), and the low frequency words in the datasets were removed.

The first dataset WebKB (available at http://www.cs.cmu.edu/˜webkb) contains six subjects (categories). There are 3,921 documents and 7,359 links. The vocabulary size is 5,019.

The second dataset Wikipedia (available at http://www.mpi-inf.mpg.de/~angelova) contains four subjects (categories): Biology, Physics, Chemistry, and Mathematics. There are 2,970 documents and 45,818 links. The vocabulary size is 3,287.

The third dataset is ODP composed of homepages of researchers and their first level outlinked pages (cited documents). We randomly selected five subjects from the ODP archive. They are Cognitive Science (CogSci), Theory, NeuralNetwork (NN), Robotics, and Statistics. There are 3,679 pages and 2,872 links. The vocabulary size is 3,529.

WebKB and Wikipedia are public datasets widely used in topic model studies. ODP was collected by us in this work.

4.2 Topic Discovery

We created four topic models HTM, LDA, Link-LDA, and Link-PLSA-LDA using all the data in each of the three datasets, and evaluated the topics obtained in the models. We heuristically set the numbers of topics as 10 for ODP, 12 for WebKB, and 8 for Wikipedia (i.e., two times of the number of true subjects). We found that overall HTM can construct more understandable topics than the other models. Figure 2 shows the topics related to the subjects created by the four models from the ODP dataset. HTM model can more accurately extract the three topics: Theory, Statistic, and NN than the other models. Both LDA and Link-LDA had mixed topics, labeled as 'Mixed' in Figure 2. Link-PLSA-LDA missed the topic of Statistics. Interestingly, all the four models split Cognitive Science into two topics (showed as CogSci-1 and CogSci-2), probably because the topic itself is diverse.

4.3 Document Classification

We applied the four models in the three datasets to document classification. Specifically, we used the word distributions of documents created by the models as feature vectors of the documents and used the subjects in the datasets as categories. We further randomly divided each dataset into three parts (training, validation, and test) and conducted 3-fold cross-validation experiments. In each trial, we trained an SVM classifier with the training data, chose parameters with the validation data, and conducted evaluation on classification with the test data. For HTM, we chose the best λ value with the validation set in each trial. Table 2 shows the classification accuracies. We can see that HTM performs better than the other models in all three datasets.

Table 2: Classification accuracies in 3-fold cross-validation.

	LDA	HTM	Link-LDA	Link-PLSA-LDA
ODP	0.640	**0.698**	0.535	0.581
WebKB	0.786	**0.795**	0.775	0.774
Wikipedia	0.845	**0.866**	0.853	0.855

Table 3: Sign-test results between HTM and the three baseline models.

	LDA	Link-LDA	Link-PLSA-LDA
ODP	0.0237	2.15e-05	0.000287
WebKB	0.0235	0.0114	0.00903
Wikipedia	1.79e-05	0.00341	0.00424

We conducted sign-tests on all the results of the datasets. In most cases HTM performs statistically significantly better than LDA, Link-LDA, and Link-PLSA-LDA (p-value < 0.05). The test results are shown in Table 3.

4.4 Discussion

We conducted analysis on the results to see why HTM can work better. Figure 3 shows an example homepage from the ODP dataset, where superscripts denote the indexes of outlinked pages. The homepage contains several topics, including Theory, Neural network, Statistics, and others, while the cited pages contain detailed information about the topics. Table 4 shows the topics identified by the four models for the homepage. We can see that HTM can really more accurately identify topics than the other models.

The major reason for the better performance by HTM seems to be that it can fully leverage the infor-

Table 4: Comparison of topics identified by the four models for the example homepage. Only topics with probabilities > 0.1 and related to the subjects are shown.

Model	Topics	Probabilities
LDA	Mixed	0.537
HTM	Theory	0.229
	NN	0.278
	Statistics	0.241
Link-LDA	Statistics	0.281
Link-PLSA-LDA	Theory	0.527
	CogSci-2	0.175

(a) LDA

Mixed	NN	Robot	CogSci-1	CogSci-2
statistic	learn	robot	visual	conscious
compute	conference	project	model	psychology
algorithm	system	file	experiment	language
theory	neural	software	change	cognitive
complex	network	code	function	experience
mathematics	model	program	response	brain
model	international	data	process	theory
science	compute	motor	data	philosophy
computation	ieee	read	move	science
problem	proceedings	start	observe	online
random	process	build	perception	mind
analysis	computation	comment	effect	concept
paper	machine	post	figure	physical
method	science	line	temporal	problem
journal	artificial	include	sensory	content

(b) HTM

Theory	Statistics	NN	Robot	CogSci-1	CogSci-2
compute	model	learn	robot	conscious	memory
science	statistic	system	project	visual	psychology
algorithm	data	network	software	experience	language
theory	experiment	neural	motor	change	science
complex	sample	conference	sensor	perception	cognitive
computation	process	model	code	move	brain
mathematics	method	compute	program	theory	human
paper	analysis	ieee	build	online	neuroscience
problem	response	international	line	physical	journal
lecture	figure	proceedings	board	concept	society
random	result	machine	read	problem	trauma
journal	temporal	process	power	philosophy	press
bound	probable	computation	type	object	learn
graph	observe	artificial	comment	content	abuse
proceedings	test	intelligence	post	view	associate

(c) Link-LDA

Statistics	Mixed	Robot	CogSci-1	CogSci-2
statistic	compute	robot	visual	conscious
model	conference	project	model	psychology
data	system	software	experiment	cognitive
analysis	learn	file	change	language
method	network	motor	function	brain
learn	computation	robotics	vision	science
sample	proceedings	informatik	process	memory
algorithm	neural	program	perception	theory
process	ieee	build	move	philosophy
bayesian	algorithm	board	response	press
application	international	sensor	temporal	online
random	science	power	object	neuroscience
distribution	complex	code	observe	journal
simulate	theory	format	sensory	human
mathematics	journal	control	figure	mind

(d) Link-PLSA-LDA

Theory	NN	Robot	CogSci-1	CogSci-2
compute	conference	robot	conscious	model
algorithm	learn	code	experience	process
computation	science	project	language	visual
theory	international	typeof	book	data
complex	system	motor	change	experiment
science	compute	control	make	function
mathematics	network	system	problem	learn
network	artificial	serve	brain	neural
paper	ieee	power	world	system
journal	intelligence	program	read	perception
proceedings	robot	software	case	represent
random	technology	file	than	vision
system	proceedings	build	mind	response
problem	machine	pagetracker	theory	object
lecture	neural	robotics	content	abstract

Figure 2: Topics identified by four models

Radford M.Neal

Professor, Dept. of Statistics and Dept. of Computer Science, University of Toronto

I'm currently highlighting the following :

- A new R function for performing univariate slice sampling.[1]
- A workshop paper on Computing Likelihood Functions for High-Energy Physics Experiments when Distributions are Defined by Simulators with Nuisance Parameters.[2]
- Slides from a talk at the Third Workshop on Monte Carlo Methods on "Short-Cut MCMC: An Alternative to Adaptation", May 2007: Postscript, PDF.

Courses I'm teaching in Fall 2008 :

- STA 437: Methods for Multivariate Data[3]
- STA 3000: Advanced Theory of Statistics[4]

You can also find information on courses I've taught in the past.[5]

You can also get to information on :

- Research interests[6] (with pointers to publications)
- Current and former graduate students[7]
- Current and former postdocs[8]
- Curriculum Vitae: PostScript, or PDF.
- Full publications list[9]
- How to contact me[10]
- Links to various places[11]

If you know what you want already, you may wish to go directly to :

- Software available on-line[12]
- Papers available on-line[13]
- Slides from talks[14]
- Miscellaneous other stuff[15]

Information in this hierarchy was last updated 2008-06-20.

Figure 3: An example homepage: http://www.cs.utoronto.ca/~ radford/

Table 5: Word assignment in the example homepage.

Word	b_{dn}	c_{dn}	Topic	Probability
mcmc	0.544	2	Stat	0.949
experiment	0.546	2	Stat	0.956
neal	0.547	8	NN	0.985
likelihood	0.550	2	Stat	0.905
sample	0.557	2	Stat	0.946
statistic	0.559	2	Stat	0.888
parameter	0.563	2	Stat	0.917
perform	0.565	2	Stat	0.908
carlo	0.568	2	Stat	0.813
monte	0.570	2	Stat	0.802
toronto	0.572	8	NN	0.969
distribution	0.578	2	Stat	0.888
slice	0.581	2	Stat	0.957
energy	0.581	13	NN	0.866
adaptation	0.591	7	Stat	0.541
teach	0.999	11	Other	0.612
current	0.999	11	Other	0.646
curriculum	0.999	11	Other	0.698
want	0.999	11	Other	0.706
highlight	0.999	10	Other	0.786
professor	0.999	11	Other	0.764
academic	0.999	11	Other	0.810
student	0.999	11	Other	0.817
contact	0.999	11	Other	0.887
graduate	0.999	11	Other	0.901

Table 6: Most salient topics in cited pages.

URL	Topic	Probability
2	Stat	0.690
7	Stat	0.467
8	NN	0.786
13	NN	0.776

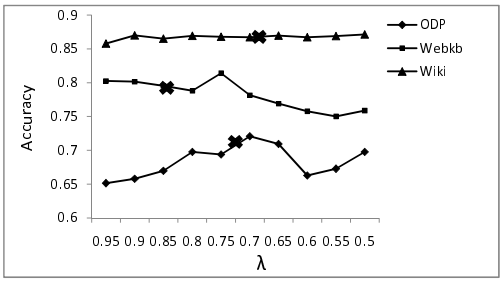

Figure 4: Classification accuracies on three datasets with different λ values. The cross marks on the curves correspond to the average values of λ in the 3-fold cross-validation experiments.

mation from the cited documents. We can see that the content of the example homepage is diverse and not very rich. It might be hard for the other baseline models to identify topics accurately. In contrast, HTM can accurately learn topics by the help of the cited documents. Specifically, if the content of a document is diverse, then words in the document are likely to be assigned into wrong topics by the existing approaches. In contrast, in HTM with propagation of topic distributions from cited documents, the words of a document can be more accurately assigned into topics. Table 5 shows the first 15 words and the last 10 words for the homepage given by HTM, in ascending order of b_{dn}, which measures the degree of influence from the cited documents on the words (the smaller the stronger). The table also gives the values of c_{dn}, indicating which cited documents have the strongest influence. Furthermore, the topics having the largest posterior probabilities for the words are also shown. We can see that the words 'experiment', 'sample', 'parameter', 'perform', and 'energy' are accurately classified. Table 6 gives the most salient topics of cited documents. It also shows the probabilities of the topics given by HTM. We can see that there is a large agreement between the most salient topics in the cited documents and the topics which are affected the most in the citing document.

Parameter λ is the only parameter in HTM which needs to be tuned. We found that the performance of HTM is not very sensitive to the values of λ, which reflects the degree of influence from the cited documents to the citing document. HTM can perform well with different λ values. Figure 4 shows the classification accuracies of HTM with respect to different λ values for the three datasets. We can see that HTM works better than the other models in most of the cases (cf., Table 2).

5 Conclusion

In this paper, we have proposed a novel topic model for hypertexts called HTM. Existing models for processing hypertexts were developed based on the assumption that both words and hyperlinks are stochastically generated by the model. The generation of latter type of data is actually unnecessary for representing *contents* of hypertexts. In the HTM model, it is assumed that the hyperlinks of hyper-

texts are given and only the words of the hypertexts are stochastically generated. Furthermore, the word distribution of a document is determined not only by the topics of the document in question but also from the topics of the documents which the document cites. It can be regarded as 'propagation' of topics reversely along hyperlinks in hypertexts, which can lead to more accurate representations than the existing models. HTM can naturally mimic human's process of creating a document (i.e., by considering using the topics of the document and at the same time the topics of the documents it cites). We also developed methods for learning and inferring an HTM model within the same framework as LDA (Latent Dirichlet Allocation). Experimental results show that the proposed HTM model outperforms the existing models of LDA, Link-LDA, and Link-PLSA-LDA on three datasets for topic discovery and document classification.

As future work, we plan to compare the HTM model with other existing models, to develop learning and inference methods for handling extremely large-scale data sets, and to combine the current method with a keyphrase extraction method for extracting keyphrases from web pages.

6 Acknowledgement

We thank Eric Xing for his valuable comments on this work.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. Modern Information Retrieval. *ACM Press / Addison-Wesley*.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of machine Learning Research*, 3:993–1022.

David Blei and John Lafferty. 2005. Correlated Topic Models. In *Advances in Neural Information Processing Systems 12*.

David Blei and John Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*.

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2007. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In *Advances in Neural Information Processing Systems 19*.

David Cohn and Huan Chang. 2000. Learning to Probabilistically Identify Authoritative Documents. In *Proceedings of the 17rd international conference on Machine learning*.

David Cohn and Thomas Hofmann. 2001. The missing link - a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems 13*.

Laura Dietz, Steffen Bickel and Tobias Scheffer. 2007. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*.

Elena Erosheva, Stephen Fienberg, and John Lafferty. 2004. Mixed-membership models of scientific publications. In *Proceedings of the National Academy of Sciences*, 101:5220–5227.

Thomas Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. In *Proceedings of the National Academy of Sciences, 101 (suppl. 1)* .

Thomas Griffiths, Mark Steyvers, David Blei, and Joshua Tenenbaum. 2005. Integrating Topics and Syntax. In *Advances in Neural Information Processing Systems, 17*.

Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. 2008. Latent Topic Models for Hypertext. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*.

Michael Jordan, Zoubin Ghahramani, Tommy Jaakkola, and Lawrence Saul. 1999. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233.

QiaoZhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic Modeling with Network Regularization. In *Proceeding of the 17th international conference on World Wide Web*.

Thomas Minka and John Lafferty. 2002. Expectation-Propagation for the Generative Aspect Model. In *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence*.

Ramesh Nallapati and William Cohen. 2008. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *International Conference for Webblogs and Social Media*.

Martin Wainwright, and Michael Jordan. 2003. Graphical models, exponential families, and variational inference. In *UC Berkeley, Dept. of Statistics, Technical Report, 2003*.

Xing Wei and Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*.