

## 基于词相似性与 CRP 的主题模型\*

张小平<sup>1</sup> 周雪忠<sup>1</sup> 黄厚宽<sup>1</sup> 冯 奇<sup>1</sup> 陈世波<sup>2</sup>

<sup>1</sup>(北京交通大学 计算机与信息技术学院 北京 100044)

<sup>2</sup>(中国中医科学院广安门医院 北京 100053)

**摘 要** 主题模型能提取隐含在文档中的主题,使文档可按主题进行归约、分类和检索,成为信息分类和检索领域的研究热点.针对 LDA(Latent Dirichlet Allocation)主题模型不能自动确定主题数目的问题,提出一种结合词相似性与 CRP(Chinese Restaurant Process)的隐主题模型,可自适应地动态更新主题内容,确定合理的主题数目.同时提出一种在动态更新主题数时超参数设置方法.在中医临床诊疗数据的实验中,获得领域专家解释性较好的分析结果.

**关键词** 主题模型,词相似性,Dirichlet 分布

中图法分类号 TP 391

## A Topic Model Based on CRP and Word Similarity

ZHANG Xiao-Ping<sup>1</sup>, ZHOU Xue-Zhong<sup>1</sup>, HUANG Hou-Kuan<sup>1</sup>, FENG Qi<sup>1</sup>, CHEN Shi-Bo<sup>2</sup>

<sup>1</sup>(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044)

<sup>2</sup>(Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100053)

### ABSTRACT

The topic model can extract the topics hid in documents to make the dimensions of documents reduced and the documents be classified and retrieved according to their topics. It is a research focus on information classification and retrieval fields. Aiming at the problem that the number of topics cannot be automatically determined in LDA topic model, a latent topic model is proposed by combining the similarity between words and Chinese restaurant process (CRP). It can adaptively update the contents and determine the rational number of topics. Meanwhile, a novel method of setting the hyperparameters during updating topics is put forward. The experimental results on traditional Chinese medicine (TCM) clinical dataset show the proposed model has good analysis results accepted by TCM expert.

**Key Words** Topic Model, Word Similarity, Dirichlet Distribution

\* 国家 973 计划项目 (No. 2006CB504601)、国家科技支撑计划项目 (No. 2007BA110B06-01)、国家自然科学基金项目 (No. 90709006) 和北京市科学技术委员会科研攻关项目 (No. D08050703020804) 资助

收稿日期:2009-04-27;修回日期:2009-10-09

作者简介 张小平,女,1969 年生,博士研究生,副教授,主要研究方向为人工智能、数据挖掘. E-mail: zh\_xping@hotmail.com. 周雪忠,男,1977 年生,博士,硕士生导师,主要研究方向为数据库、数据挖掘、医学本体论与中医信息学. 黄厚宽,男,1940 年生,教授,博士生导师,主要研究方向为人工智能、数据挖掘、机器学习. 冯奇,男,1982 年生,博士研究生,主要研究方向为数据挖掘、POMDP. 陈世波,男,1973 年生,博士,主治医师,主要研究方向为糖尿病及其并发症的中医药防治研究、个体化诊疗及临床评价.

## 1 引言

主题模型(Topic Model)<sup>[1-7]</sup>的基本思想是假设存在  $K$  个隐主题,其中每个主题是词的多项式分布,而文档是由这  $K$  个隐主题随机混合产生.主题模型指出文档的简单概率生成过程,它是一种产生式概率模型(Generative Probabilistic Model).

主题模型是在 N-gram 模型<sup>[8]</sup>、混合 unigram 模型(Mixture of Unigram)<sup>[9]</sup>、隐语义索引(Latent Semantic Indexing, LSI)<sup>[6]</sup>以及概率隐语义索引(Probabilistic Latent Semantic Indexing, PLSI)<sup>[7]</sup>等一系列隐模型背景下产生的. N-gram 模型<sup>[8]</sup>是概率词频模型,每个文档的词是从一个多项式分布中独立抽取出来的.它没有考虑文本的主题.混合 Unigrams 模型<sup>[9]</sup>通过在 Unigrams 模型中增加一个离散的随机主题变量  $z$  而获得,该模型假设每个文档只由一个主题生成.实际上,一篇文档可由多个主题组成.概率隐语义索引模型 PLSI<sup>[7]</sup>放松了混合 Unigrams 模型中每个文档只有一个主题假设,允许存在多个主题.然而,由于学习出的主题直接依赖于训练文档,导致主题的参数个数随训练文档集中文档数的增加而线性增加,模型会产生过度拟合,不适合预测新的数据. LDA (Latent Dirichlet Allocation) 主题模型<sup>[11]</sup>通过引入 2 个 Dirichlet 先验参数,解决 PLSI 模型参数个数线性增加和过度拟合的问题.

许多模型对 LDA 进行改进和扩展.分层主题模型 hLDA (Hierarchical LDA)<sup>[10]</sup>、PAM (Pachinko Allocation Model)<sup>[11]</sup>模型可以获得主题间的层次关系. CTM (Correlated Topic Model) 模型<sup>[12-13]</sup>则可解决主题间的相关性问题.有监督的主题模型 sLDA (Supervised LDA)<sup>[14]</sup>通过引入人类标签,使得建立和预测的主题更精确.作者-主题模型 (Author-Topic Model)<sup>[15]</sup>分析文档主题时引入了作者信息.主题模型已经成为信息检索<sup>[16]</sup>、文本挖掘<sup>[17-19]</sup>、社会网<sup>[20]</sup>、生物信息学<sup>[21]</sup>等领域的研究热点.而模型参数问题,如最优主题数、超参数设置是研究难点.

LDA 模型的主题数是由人工确定<sup>[1]</sup>或通过反复多次实验确定符合一定判断标准的主题数<sup>[4,18]</sup>.本文基于 CRP 自动确定相互独立的主题,通过引入词相似性信息快速确定词所属的主题.模拟数据实验和中医临床诊疗数据实验结果表明,由于参照词间的相似性,经过少数几次迭代后,算法就能收敛,获得解释性较好的主题.

## 2 LDA 和 CRP

### 2.1 LDA

LDA 模型是一个具有 3 层的产生式 Bayesian 概率模型(如图 1 所示),由文档、主题和词组成.

假设文档集有  $D$  个文档,每个文档看作是由  $K$  个独立的主题混合产生,其中  $K$  假设已知并且相互独立,每个主题  $k$  由词上的多项式分布形成.图中词  $w$  用灰色标志,它是唯一能够观察到的变量.  $w_{dn}$  表示第  $d$  个文档的第  $n$  个词,  $w_{dn} \in V$ ,  $V$  是词的字典集.  $z_{dn}$  表示产生  $w_{dn}$  的主题.  $\alpha$  是文档集的先验分布参数.  $\theta_d$  是文档  $d$  在主题上的分布,对于每个文档  $d$ ,多项式分布  $\theta_d$  的提取服从 Dirichlet 分布  $Dir(\theta_d | \alpha)$ . 一个主题  $\beta_k$  是字典  $V$  中词上的分布.图中模型包含  $K$  个主题  $\beta_{1:K}$ ,  $N$  是文档  $d$  的总词数.给定参数  $\alpha, \beta$ , 文档  $d$  的隐变量  $\theta_d, z_d$  和显变量  $w_d$  的联合分布是

$$p(w_d, z_d, \theta_d | \alpha, \beta) = p(\theta_d | \alpha) \prod_{i=1}^N p(z_{di} | \theta_d) p(w_{di} | z_{di}, \beta).$$

由于同时存在多个隐变量,上式不能直接通过推理求出.文献[1]采用变分 Bayes 推理(Variational Bayes Inference)算法进行推理隐变量.文献[4]、[22]采用 Gibbs 采样方法估计当前采样词  $w_{di}$  的主题  $z_{di}$  的后验分布,然后得到模型参数  $\theta$  和  $\phi$ ,其中  $\phi_{mj}$  是词  $m$  在主题  $j$  上的概率,  $\theta_{dj}$  是文档  $d$  运用主题  $j$  的概率.其基本思想是利用已采样得到的所有主题在词上的分布  $\phi$  和所有文档在主题上的分布  $\theta$  推断当前采样的词  $w_{di}$  所属的主题  $z_{di}$ ,即

$$p(z_{di} = j | w_{di} = m, z_{-i}, w_{-i}) \propto \frac{C_{mj}^{WT} + \beta}{\sum_m C_{mj}^{WT} + V\beta} \frac{C_{dj}^{DT} + \alpha}{\sum_j C_{dj}^{DT} + T\alpha},$$

$$\phi_{mj} = \frac{C_{mj}^{WT} + \beta}{\sum_m C_{mj}^{WT} + V\beta}, \quad \theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_j C_{dj}^{DT} + T\alpha}, \quad (1)$$

其中,  $C_{mj}^{WT}$ 、 $C_{dj}^{DT}$  分别表示词  $w_{di}$  的值  $m$  在主题  $j$  上出现的个数和文档  $d$  出现主题  $j$  上的个数,  $z_{di} = j$  表示分配文档  $d$  的第  $i$  个词的主题为  $j$ .

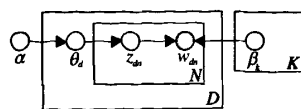


图1 隐 Dirichlet 分配主题模型

Fig. 1 Latent Dirichlet allocation topic model

## 2.2 CRP

CRP<sup>[23]</sup> 是分配  $M$  个客人到某餐馆的餐桌上就坐的离散过程,其中假设餐馆有无限张餐桌,它是一种 Dirichlet 过程. 分配过程如下:第 1 个客人坐在第 1 张餐桌,第  $m$  个客人所选餐桌  $z_m = j$  的概率与桌子  $j$  上的客人的分布有关:

$$p(z_m = j | z_{-m}) = \begin{cases} \frac{C_j}{m-1+\gamma}, & j \text{ 餐桌有客人 } C_j \text{ 个} \\ \frac{\gamma}{m-1+\gamma}, & j \text{ 餐桌没有客人} \end{cases}$$

其中  $\gamma$  是 Dirichlet 先验参数.

CRP 分配客人到餐桌上的过程只考虑餐桌上客人的个数,没有考虑将要就坐的客人餐桌上的客人相关性. 假设某张餐桌人气很旺,但都是将要就坐的客人不喜欢的人,他也不会选择这张餐桌,这也符合日常习惯.

## 3 基于词相似性和 CRP 的自动生成主题方法

该算法考虑了餐桌上的客人对将要就坐的客人的吸引力. 对于 LDA 主题,如果把一张餐桌看作是一个主题,餐桌上的客人看作是主题上的词,则 Gibbs 采样的过程如下.

词  $m$  分配到已有的  $T-1$  个主题中的第  $j$  个主题的概率如下:

$$p(z_{di} = j | w_{di} = m, z_{-i}, w_{-i}, \text{sim}) \propto \frac{\sum_{m' \setminus m} (C_{m'j}^{WT} \cdot \text{sim}_{m'm}) + C_{mj}^{WT} + \beta}{\sum_{m'} (C_{m'j}^{WT} + V\beta)} \cdot \frac{C_{dj}^{DT} + \alpha}{\sum_j (C_{dj}^{DT} + T\alpha)}, \quad (2)$$

其中  $\text{sim}$  是已知词间的稀疏相似矩阵.

$$\phi_{mj} = \frac{\sum_{m' \setminus m} (C_{m'j}^{WT} \cdot \text{sim}_{m'm}) + C_{mj}^{WT} + \beta}{\sum_{m'} (C_{m'j}^{WT} + V\beta)}, \quad (3)$$

$$\theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_j (C_{dj}^{DT} + T\alpha)}, \quad (4)$$

词  $m$  分配到一个新主题  $j = T$  概率如下

$$p(z_{di} = j | w_{di} = m, z_{-i}, w_{-i}) \propto \frac{0 + \beta}{0 + V\beta} \cdot \frac{0 + \alpha}{\sum_j (C_{dj}^{DT} + T\alpha)}. \quad (5)$$

$\sum_{m' \setminus m} (C_{m'j}^{WT} \cdot \text{sim}_{m'm})$  表示主题  $j$  中除了词  $m$  的其它所有词对  $m$  形成的吸引力,其中  $\text{sim}_{m'm}$  表示词  $m$

与  $m'$  的相似度,  $\sum_{m' \setminus m} (C_{m'j}^{WT} \cdot \text{sim}_{m'm}) + C_{mj}^{WT}$  表示主题  $j$  对词  $m$  形成的总吸引力. 若稀疏相似矩阵中包含相同词间的相似度为 1 这个值,则公式

$$\sum_{m' \setminus m} (C_{m'j}^{WT} \cdot \text{sim}_{m'm}) + C_{mj}^{WT}$$

改写为  $\sum_{m'} (C_{m'j}^{WT} \cdot \text{sim}_{m'm})$ .

具体算法如下.

step 1 分配训练集的第一个词  $w_1$  的主题为 1; 利用式(3)、(4) 更新主题 1 在  $w_1$  的分布  $\phi$  以及  $w_1$  所在的文档在主题 1 分布  $\theta$ .

step 2 repeat

for 训练集除  $w_1$  以外的  $n-1$  个词  $w_i$

根据式(2)和(5) 分配  $w_i$  到一个已存在主题或一个新主题,然后利用式(3)、(4) 更新  $\phi$  以及  $\theta$ ;

If 分配词  $w_i$  到一个新主题,主题数累加 1

end for

until 达到最大迭代次数或收敛

## 4 超参数设置

在主题模型中,超参数设置对分析结果具有关键作用,然而,目前仍处于经验阶段. 在一些相关研究<sup>[4,19,22]</sup> 中提到 Dirichlet 超参数  $\alpha, \beta$  的设置方法是令  $\alpha = 50/T, \beta = 0.01$  或  $\beta = 0.1$ . 该设置方法不能灵活应用于各种主题模型中.

本文提出一种  $\alpha, \beta$  超参数设置方法. 根据式(1) 知  $V\beta$  是矩阵  $V \times T$  加在  $j$  列上的伪参数,即  $\beta$  是加在矩阵  $V \times T$  第  $j$  列上每个点的伪参数,或叫做平滑参数. 若  $\beta$  值为 0,则得出的模型完全依赖训练数据,存在过度拟合现象. 假设第  $j$  列平滑参数和为  $\alpha$ ,则每个  $\beta$  的值为  $\alpha/V$ . 同理,  $T\alpha$  是矩阵  $D \times T$  加在  $d$  行上的伪参数,可以设  $\alpha = b/T, a, b$  的值分别是  $V \times T$  矩阵第  $j$  列总词数以及  $D \times T$  矩阵第  $d$  行总词数的伪系数比例. 实验中设置  $a = 0.1n/T, b = 0.1n/D$ .

## 5 实验及分析

分别在一个人工模拟数据集(表 1 所示)和中医科学院数据仓库<sup>[24]</sup> 中筛选导出的 II 型糖尿病临床诊疗数据集上,利用 PC 机(C2D E8400, 3GHz, 2GB 内存),在 Matlab 运行环境中,各进行 10 次实验,计算 log 似然.

$$\log likelihood(D_{\text{train}}) = \sum_{d=1}^M \log p(d_i)$$

的平均值和语言模型复杂度  $perplexity^{[1]}$ :

$$perplexity(D_{\text{test}}) = \exp\left(-\frac{\sum_{i=1}^M \log(d_i)}{\sum_{i=1}^M N_i}\right)$$
$$\propto \exp\left(-\frac{\sum_{i=1}^M \log\left(\sum_{j=1}^{N_i} \sum_{k=1}^T p(w_{ij})\right)}{\sum_{i=1}^M N_i}\right)$$

的平均值,并与 LDA 算法<sup>[1]</sup> 在运行速度和合理的主  
题数方面进行比较.

5.1 模拟数据集

首先通过

$$Similarity(w_i, w_j) = \frac{|c|}{|a+b-c|}$$

(6)

计算表 1 中词之间的相似度,其中  $c$  表示词  $w_i$  和  $w_j$   
共同出现的文档数, $a$  表示词  $w_i$  出现的文档数, $b$  表  
示词  $w_j$  出现的文档数. 这里只取相似度大于等于  
0.1 的稀疏相似矩阵.

表 1 模拟数据集

Table 1 Simulation dataset

A	B	C	A	B	C	A	B	C
1	1	1	5	1	2	4	2	1
1	2	1	5	8	2	4	3	1
1	3	1	5	9	2	4	4	1
1	4	1	6	1	2	4	5	1
1	5	1	6	7	2	4	6	1
1	6	1	6	8	2	8	1	2
2	2	1	6	9	2	8	7	2
2	3	1	7	1	1	8	8	2
2	4	1	7	2	1	8	9	2
2	5	1	7	3	1	3	4	1
3	2	1	7	4	1	3	5	1
3	3	1	3	6	1			

运行本文提出的算法 10 次,设置  $\alpha =$   
 $0.1n/(TD)$ ,  $\beta = 0.1n/(VT)$ ,平均迭代2.9次收敛,  
平均运行时间为 0.032 7s,平均 log 似然为  
-42.274 8,主题数  $T$  主要在 2 ~ 3 间波动,表 1 给出  
模拟训练数据集在主题数为 2 的一次运行结果,表  
中, $A$  列表示文档标号, $B$  列表示词标号, $C$  列的值为  
获得的主题标号.

若不考虑词间相似性只考虑自动生成主题的  
CRP 算法时,运行 10 次,平均迭代 9.3 次收敛,平均

运行时间为 0.038 8s,平均 log 似然为 -43.688 3,主  
题数也主要在 2 ~ 3 间波动.

而运行 LDA 时,令主题  $T$  循环,  $\alpha$  和  $\beta$  取不同  
值,各运行 LDA 算法 10 次,各迭代 70 次,平均 log  
似然结果如图 2 所示.可以看出  $T = 2$  且  $\alpha =$   
 $0.1n/(TD)$ ,  $\beta = 0.1n/(VT)$  时,log 似然值最大,其  
中  $n$  为训练集的词总数.主题数从 2 ~ 10 循环运行  
一轮,所需运行时间平均为 0.397 755s.

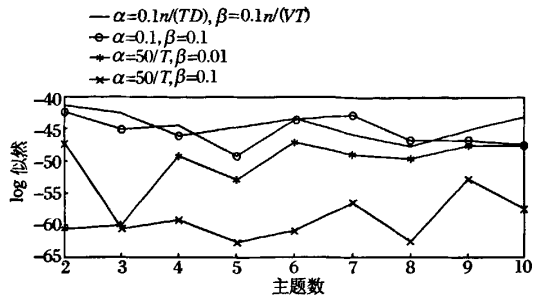


图 2 LDA 不同超参数的 log 似然值  
Fig. 2 Log likelihood for different super-parameters of LDA

5.2 中医临床糖尿病诊疗数据集

取 3 238 个糖尿病患者的中医诊疗数据,427 个  
症状,症状出现次数  $n = 39\,770$ . 首先按照式(6)  
计算症状之间的相似度,只取相似度大于等于 0.1 的  
稀疏相似矩阵,然后去掉前 15 个症状频率最高的症  
状,这些就如文本中的停用词.

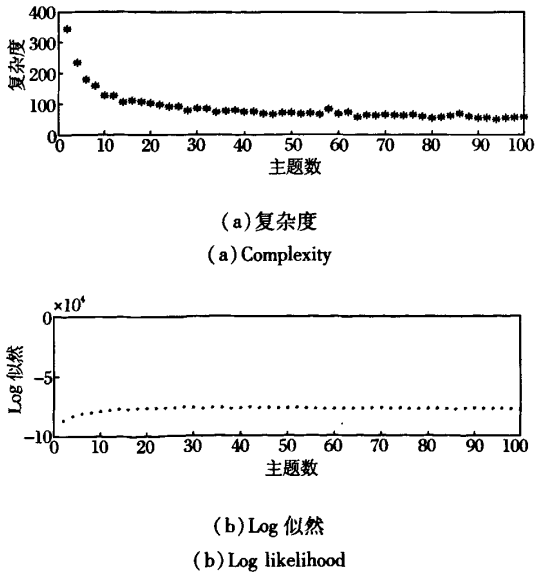


图 3 临床数据集的 log 似然及复杂度  
Fig. 3 Log likelihood and complexity on clinic dataset

采用 10-折交叉验证,运行本文算法 10 次,平均迭代 3.6 次算法收敛,主题数主要在 40 ~ 60 间波动,平均 perplexity 值为 36.574 0,平均 log 似然为 -55 120. 运行 LDA 算法 10 次,主题数从 2 ~ 100 变动,迭代次数设置为 500,平均 perplexity 值及平均 log 似然如图 3 所示. 可以看出本文算法能自动找到 perplexity 较小、log 似然较大的合理主题数. 且迭代 3.6 次本文算法所花费的时间平均为 50.977 845s,远小于设置主题数为 2 ~ 50、步长为 2、循环运行 LDA 至迭代收敛所需的平均运行时间 525.450 995s.

## 6 结束语

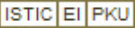
本文提出一种自动生成隐主题的方法. 利用 CRP 可自动确定主题个数,而结合已知的词间的相似关系可快速确定词对应的主题,避免同一词在不同主题间波动. 从上述实验得知,本文算法能够较快地找到合理主题数,且在实际领域数据的分析中能得到较好的结果. 通过反复实验,发现超参数对结果的影响较大,本文提出一种自动设置超参数方法. 今后将对多关系主题模型和分层多关系主题模型结合词的相似程度进行研究,并把他们应用在中医临床诊疗等实际数据中,自动发现并确认有临床意义的主题知识.

## 参 考 文 献

- [1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003, 3: 993 - 1022
- [2] Griffiths T L, Steyvers M. A Probabilistic Approach to Semantic Representation // *Proc of the 24th Annual Conference of the Cognitive Science Society*. Fairfax, USA, 2002: 381 - 386
- [3] Griffiths T L, Steyvers M. Prediction and Semantic Association // Becker S, Thrun S, Obermayer K, eds. *Advance in Neural Information Processing Systems*. Cambridge, USA: MIT Press, 2003, 15: 11 - 18
- [4] Griffiths T L, Steyvers M. Finding Scientific Topics. *Proc of the National Academy of Science*, 2004, 101 (21): 5228 - 5235
- [5] Hofmann T. Probabilistic Latent Semantic Analysis // *Proc of the 15th Conference on Uncertainty in Artificial Intelligence*. Stockholm, Sweden, 1999: 289 - 296
- [6] Hofmann T. Probabilistic Latent Semantic Indexing // *Proc of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Berkeley, USA, 1999: 50 - 57
- [7] Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*. 2001, 42 (1/2): 177 - 196
- [8] Banerjee S, Pedersen T. The Design, Implementation and Use of the Ngram Statistics Package // *Proc of the 4th International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico, Mexico, 2003: 370 - 381
- [9] Nigam K, McCallum A, Thrun S. *et al.* Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learning*, 2000, 39 (2/3): 103 - 134
- [10] Blei D, Griffiths T, Jordan M, *et al.* Hierarchical Topic Models and the Nested Chinese Restaurant Process // Thrun S, Saul L K, Schölkopf B, eds. *Advances in Neural Information Processing Systems*. Cambridge, USA: MIT Press, 2004, 16: 17 - 24
- [11] Li Wei, McCallum A. Pachinko Allocation: DAG - Structured Mixture Models of Topic Correlations // *Proc of the 23rd International Conference on Machine Learning*. New York, USA, 2006: 577 - 584
- [12] Blei D, Lafferty J. Correlated Topic Models // Weiss Y, Schölkopf B, Platt J, eds. *Advances in Neural Information Processing Systems*. Cambridge, USA: MIT Press, 2006, 18: 147 - 154
- [13] Blei D, Lafferty J. A Correlated Topic Model Science. *The Annals of Applied Statistics*, 2007, 1 (1): 17 - 35
- [14] Blei D, McAuliffe J. Supervised Topic Models // Platt J C, Koller D, Singer Y, eds. *Advances in Neural Information Processing Systems*. Cambridge, USA: MIT Press, 2008, 20: 121 - 128
- [15] Rosen-Zvi M, Griffiths T, Steyvers M, *et al.* The Author-Topic Model for Authors and Documents // *Proc of the 20th Conference on Uncertainty in Artificial Intelligence*. Banff, Canada, 2004: 487 - 494
- [16] Bhattacharya I, Getoor I. A Latent Dirichlet Model for Unsupervised Entity Resolution // *Proc of the International Conference on Data Mining*. New York, USA, 2006: 47 - 58
- [17] Li Wenbo, Sun Le, Zhang Dakun. Text Classification Based on Labeled-LDA Model. *Chinese Journal of Computers*, 2008, 31 (4): 620 - 627 (in Chinese)  
(李文波, 孙乐, 张大鲲. 基于 Labeled-LDA 模型的文本分类新算法. *计算机学报*, 2008, 31 (4): 620 - 627)
- [18] Cao Juan, Zhang Yongdong, Li Jintao, *et al.* A Method of Adaptively Selecting Best LDA Model Based on Density. *Chinese Journal of Computers*, 2008, 31 (10): 1780 - 1787 (in Chinese)  
(曹娟, 张勇东, 李锦涛, 等. 一种基于密度的自适应最优 LDA 模型选择方法. *计算机学报*, 2008, 31 (10): 1780 - 1787)
- [19] Shi Jin, Hu Ming, Shi Xin, *et al.* Text Segmentation Based on Model LDA. *Chinese Journal of Computers*, 2008, 31 (10): 1865 - 1873 (in Chinese)  
(石晶, 胡明, 石鑫, 等. 基于 LDA 模型的文本分割. *计算机学报*, 2008, 31 (10): 1865 - 1873)
- [20] McCallum A, Corrada-Emmanuel A, Wang Xuerui. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. *Journal of Artificial Intelligence Research*, 2007, 30: 249 - 272
- [21] Flaherty P, Gaever G, Kumm J, *et al.* A Latent Variable Model for Chemogenomic Profiling. *Bioinformatics*, 2005, 21 (15): 3286 - 3293
- [22] Steyvers M, Griffiths T. Probabilistic Topic Models // Landauer T, McNamara D, Dennis S, *et al.*, eds. *Handbook of Latent Semantic Analysis*. Hillsdale, USA: Erlbaum, 2007: 427 - 448
- [23] Aldous D. *Exchangeability and Related Topics*. Berlin, Germany: Springer Press, 1985: 1 - 198
- [24] Zhou Xuezhong, Liu Baoyan, Wang Yinghui, *et al.* Building Clinical Data Warehouse for Traditional Chinese Medicine Knowledge Discovery // *Proc of the International Conference on BioMedical Engineering and Informatics*. Sanya, China, 2008: 615 - 620

作者：[张小平](#)，[周雪忠](#)，[黄厚宽](#)，[冯奇](#)，[陈世波](#)，[ZHANG Xiao-Ping](#)，[ZHOU Xue-Zhong](#)，[HUANG Hou-Kuan](#)，[FENG Qi](#)，[CHEN Shi-Bo](#)

作者单位：[张小平](#)，[周雪忠](#)，[黄厚宽](#)，[冯奇](#)，[ZHANG Xiao-Ping](#)，[ZHOU Xue-Zhong](#)，[HUANG Hou-Kuan](#)，[FENG Qi](#) (北京交通大学, 计算机与信息技术学院, 北京, 100044)，[陈世波](#)，[CHEN Shi-Bo](#) (中国中医科学院广安门医院, 北京, 100053)

刊名：[模式识别与人工智能](#) 

英文刊名：[PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE](#)

年，卷(期)：2010, 23(1)

被引用次数：1次

参考文献(24条)

1. [Blei D M](#), [Ng A Y](#), [Jordan M I](#) [Latent Dirichlet Allocation](#) [外文期刊] 2003
2. [Griffiths T L](#), [Steyvers M](#) [A Probabilistic Approach to Semantic Representation](#) 2002
3. [Griffiths T L](#), [Steyvers M](#) [Prediction and Semantic Association](#) [外文会议] 2003
4. [Griffiths T L](#), [Steyvers M](#) [Finding Scientific Topics](#) [外文期刊] 2004(z1)
5. [Hofmann T](#) [Probabilistic Latent Semantic Analysis](#) 1999
6. [Hofmann T](#) [Probabilistic Latent Semantic Indexing](#) [外文会议] 1999
7. [Hofmann T](#) [Unsupervised Learning by Probabilistic Latent Semantic Analysis](#) [外文期刊] 2001(1/2)
8. [Banerjee S](#), [Pedersen T](#) [The Design, Implementation and Use of the Ngram Statistics Package](#) 2003
9. [Nigam K](#), [McCallum A](#), [Thrun S](#) [Text Classification from Labeled and Unlabeled Documents Using EM](#) [外文期刊] 2000(2/3)
10. [Blei D](#), [Griffiths T](#), [Jordan M](#) [Hierarchical Topic Models and the Nested Chinese Restaurant Process](#) [外文会议] 2004
11. [Li Wei](#), [McCallum A](#) [Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations](#) 2006
12. [Blei D](#), [Lafferty J](#) [Correlated Topic Models](#) [外文会议] 2006
13. [Blei D](#), [Lafferty J](#) [A Correlated Topic Model Science](#) 2007(01)
14. [Blei D](#), [McAuliffe J](#) [Supervised Topic Models](#) 2008
15. [Rosen-Zvi M](#), [Griffiths T](#), [Steyvers M](#) [The Author-Topic Model for Authors and Documents](#) [外文会议] 2004
16. [Bhattacharya I](#), [Getoor I](#) [A Latent Dirichlet Model for Unsupervised Entity Resolution](#) 2006
17. [李文波](#), [孙乐](#), [张大鲲](#) [基于Labeled-LDA模型的文本分类新算法](#) [期刊论文] - [计算机学报](#) 2008(04)
18. [曹娟](#), [张勇东](#), [李锦涛](#) [一种基于密度的自适应最优LDA模型选择方法](#) [期刊论文] - [计算机学报](#) 2008(10)
19. [石晶](#), [胡明](#), [石鑫](#) [基于IDA模型的文本分割](#) [期刊论文] - [计算机学报](#) 2008(10)
20. [McCallum A](#), [Corrada-Emmanuel A](#), [Wang Xuerui](#) [Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email](#) 2007
21. [Flaherty P](#), [Giaever G](#), [Kumm J](#) [A Latent Variable Model for Chemogenomic Profiling](#) [外文期刊] 2005(15)
22. [Steyvers M](#), [Griffiths T](#) [Probabilistic Topic Models](#) 2007
23. [Aldous D](#) [Exchangeability and Belated Topics](#) 1985
24. [Zhou Xuezhong](#), [Liu Baoyan](#), [Wang Yinghui](#) [Building Clinical Data Warehouse for Traditional Chinese Medicine Knowledge Discovery](#) 2008

#### 本文读者也读过(2条)

1. [林洋港](#) [概率主题模型在文本分类中的应用研究](#)[学位论文]2009
2. [张小平](#). [周雪忠](#). [黄厚宽](#). [冯奇](#). [陈世波](#). [焦宏官](#). [ZHANG Xiaoping](#). [ZHOU Xuezhong](#). [HUANG Houkuan](#). [FENG Qi](#). [CHEN Shibo](#). [JIAO Hongguan](#) [一种改进的LDA主题模型](#)[期刊论文]-[北京交通大学学报](#)2010, 34(2)

#### 引证文献(1条)

1. [曾嘉](#). [严建峰](#). [龚声蓉](#) [复杂文本网数据的主题建模进展](#)[期刊论文]-[计算机学报](#) 2012(12)

本文链接: [http://d.wanfangdata.com.cn/Periodical\\_mssbyrgzn201001012.aspx](http://d.wanfangdata.com.cn/Periodical_mssbyrgzn201001012.aspx)