基于概率主题模型的文档聚类

王李冬[1,2],魏宝刚[1],袁 杰[1]
(1. 浙江大学计算机科学与技术学院,浙江杭州 310027;2. 杭州师范大学,浙江杭州 310012)

摘 要: 为了实现普通文本语料库和数字图书语料库的有效聚类,分别提出基于传统 LDA(Latent Dirichlet Allocation)模型和 TC_LDA 模型的聚类算法.TC_LDA 模型在 LDA 模型基础上进行扩展,通过对图书文档的目录和正文信息联合进行主题建模.和传统方法不同,基于主题模型的聚类算法能将具备同一主题的文档聚为一类.实验结果表明从主题分析角度出发实现的聚类算法优于传统的聚类算法.

关键词: 主题模型; LDA 模型; TC_LDA 模型; 文档聚类
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2012)11-2346-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.11.033

Document Clustering Based on Probabilistic Topic Model

WANG Li-dong[1,2], WEI Bao-gang[1], YUAN Jie[1]
(1. *College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*;
2. *Hangzhou Normal University, Hangzhou, Zhejiang* 310012, *China*)

Abstract: To effectively cluster corpus of ordinary documents and digital books, the clustering algorithms based on LDA model and TC_LDA were proposed, respectively. The topic model named TC_LDA, the extension of LDA, is proposed for digital books corpus for jointly topic modeling from both of Texts and Contents. Unlike traditional clustering methods, topic model based methods cluster documents in a group if they share one or more common topics. Empirical evaluation demonstrates that our approach based on topic analysis can substantially improve the clustering results as compared to related methods.

Key words: topic model; LDA model; TC_LDA model; document clustering

1 引言

基于因特网和数字图书馆的庞大数字资源为深层次的数据挖掘提供了一定的资源条件.聚类作为一种传统的数据挖掘技术和非监督学习方法[1],试图将文档分割为不同的类别,且同一类别内的所有文档能够共享同一主题.然而,大多数传统的聚类方法存在一定的弊端,忽略了聚类问题的一个本质即主题偏差,仅仅依照词汇的共现规则无法解释两篇文档是否具备共同主题.目前也有很多工作针对大文本聚类集中于降维技术研究[2],但是该技术除了存在信息失真外,当文档存在大量的非相关属性时,降维并不能提升聚类效果.因此,若针对大文本聚类,传统方法不能取得满意的效果.

本文研究的主要目的在于提出一种聚类算法能有效应用于大文本和普通文本.基于此,本文提出了一种新的通过主题分析实现的有效聚类算法.该方法从主题角度出发,使得聚为同类的文档含有相似的一个或多个主题.针对普通文档,提出一种基于 LDA 模型的聚类方法;针对图书文档(大文本),利用数字图书的多粒度信息(包括目录以及章节正文),通过对正文信息的特征过滤,对主题模型 LDA 进行改进得到 TC_LDA 模型,有效的提取目录和章节正文的主题信息,再根据各粒度主题信息中词汇的分布,得到不同图书的三层分布模型,即图书—主题—词汇概率分布结构.最终,通过对普通文档语料库和数字图书文档语料库的实验,证明该方法比其他的传统聚类方法效果更好.

2 基于 LDA 的聚类算法

LDA 模型在文本相关应用领域中已经获得较好的效果,如文本分类[3],信息检索[4,5]等.LDA 模型能有效解决概率潜在语义分析(PLSA)模型的过度拟合问题.LDA 模型属于生成模型(generative modal),文本和词汇间的关系通过隐藏变量 z_n 体现,z_n 代表文档中的某一特定主题.文档和词汇满足 Dirichlet 分布,α,β 分别为

收稿日期:2011-01-18;修回日期:2012-05-10
基金项目:国家自然科学青年基金(No.61103171,No.61103099);浙江省公益性技术应用研究计划(No.2011C31048)

相应 Dirichlet 分布的参数.

LDA 模型的核心问题是隐含变量的概率分布的估计,即获得目标文档中的隐含主题分布 $\theta^{d_i}_{z=j}$以及各隐含主题中的词汇分布 $\varphi^{z=j}_{w_i}$(具体实现过程可见文献[6]中的 Gibbs 采样).经过足够次数的采样迭代,$\theta^{d_i}_{z=j}$可以通过下列计算表达式进行估算:

$$\theta^{d_i}_{z=j} = p(z=j \mid d_i) = \frac{n^{d_i}_{z=j} + \alpha}{n^{d_i}_{\cdot} + K\alpha} \tag{1}$$

其中,K 代表主题个数,N 代表文档中的词汇个数.$n^{d_i}_{z=j}$代表文档 d_i 中分配给主题 j 的词汇个数,$n^{d_i}_{\cdot}$代表 d_i 的词汇个数.在隐含主题分布 $\theta^{d_i}_{z=j}$的基础之上,再根据主题建模得到的文档-主题概率分布结构,定义主题相关度函数计算两篇文档是否同时包含同一主题,并由此判定两篇文档是否相关.

定义 1(文档的主题相关度计算函数)　文档集合中任意两篇文档 d_i, d_j 之间的主题相关度函数定义为:

$$Corr(d_i, d_j) = \begin{cases} 1, \text{if } \forall\, n \in [1\cdots K], \exists\,(p(z=n \mid d_i) > T \wedge p(z=n \mid d_j) > T) \\ 0, \quad \text{else} \end{cases} \tag{2}$$

上式 K 代表主题个数,T 为判定阈值.针对 $\forall\, d_i, d_j \in D$,进行上式的计算,便可根据以下规则判定文档 d_i, d_j 是否聚为同类:(a)若 $Corr(d_i, d_j) = 1$,则 d_i, d_j 聚为同类;(b)若 $Corr(d_i, d_j) = 1$ 且 $Corr(d_j, d_k) = 1$,则 d_i, d_j, d_k 聚为同类.为了进一步简化计算,分析公式(1)可得,给定一篇文档 d_i,其主题 z 被赋予 n 的概率 $p(z = n \mid d_i)$是由分量 $n^{d_i}_{z=n}$决定的.于是,我们直接计算 $n^{d_i}_{z=n}$的值,再根据相应的聚类效果设定阈值 T.

3　基于 TC_LDA 模型的聚类

3.1　TC_LDA 模型简介

数字图书属于非结构化信息,除了书名之外,包含目录、正文等信息,部分图书含有摘要.数字图书的目录本身含有丰富的主题信息,但是由于它所包含的信息较精简,用它单独代表整本书进行特征量提取往往存在很大的失真.其次,若直接将目录和正文整合在一起作为一篇独立的正文*进行主题提取,则会将目录信息当作普通的正文信息来处理,会隐盖目录的重要性.鉴于上述因素,本文提出了一种联合目录和正文信息进行主题建模的聚类方法,该方法在 LDA 模型的基础上扩展,将目录和正文作为独立的可观察信息进行主题建模,并将其整合到同一模型中,称为 TC_LDA 模型,其贝叶斯网络图如图 1 所示.

图 1 中菱形区域里的值表示参数,D 表示图书文档的个数,N 表示正文信息词汇的个数,来自于长度为 $|T|$的词汇表,M 表示目录信息词汇的个数,来自长度为 $|C|$的词汇表.K 和 L 分别表示正文和目录的主题个数.w_t, w_c 的所在圆圈被填充表示该随机变量为观察变量,未被填充的圆圈内为隐含变量.

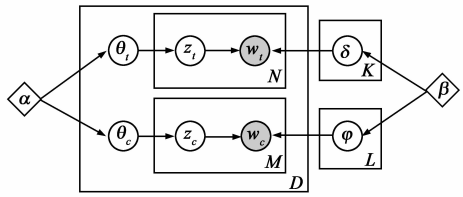

图1　TC_LDA模型的概率模型图

该模型生成一册图书需具备以下步骤:(1)针对每一册图书 $i \in [i \cdots D]$,分别选择 θ_{ti}和 θ_{ci},服从 *Dirichlet*(α)分布.θ_{ti}代表给定某一图书文档 i,主题为 k 的概率,针对文档中的正文信息,$k \in [1 \cdots K]$.θ_{ci}代表给定某一图书文档 i,主题为 l 的概率,针对文档中的目录信息,$l \in [1 \cdots L]$.(2)针对正文信息,选择 δ_K,长度为 $|T|$,服从 *Dirichlet*(β)分布.δ_K 代表给定主题 k,能观察到所有词汇的概率.(3)针对目录信息,选择 φ_l,长度为 $|C|$,服从 *Dirichlet*(β)分布.φ_l 代表给定主题 l,能观察到所有词汇的概率.(4)对文档 i 中正文内容的每一词汇 $t \in [1 \cdots N]$:(a)根据 θ_{ti},选择主题 $z_t, z_t \in [1 \cdots K]$.($b$)根据 δ_{z_t},选择词汇 w_t.(5)对文档 i 中目录内容的每一词汇 $c \in [1 \cdots M]$:(a)根据 θ_{ci},选择主题 $z_c, z_c \in [1 \cdots L]$.($b$)根据 φ_{z_c},选择词汇 w_c.

由上面步骤可得,TC_LDA 模型中针对目录信息的主题建模方法类似于正文信息的主题建模.正文信息的主题数目和目录信息的主题数目需设为不同的值,这是因为目录信息和正文信息所包含的信息量不同所决定的,正文提取出来的主题数应远大于目录信息的主题数.

3.2　TC_LDA 参数学习

该部分同样采用 Gibbs 抽样[6]对参数 $\theta_t, \theta_c, \delta, \varphi$ 进行学习,并将图书文档的目录和正文分割为独立的文档实现参数估计.经过足够次的迭代(迭代次数预先设定),$\theta_t, \theta_c, \delta, \varphi$ 的近似估计见公式(3)~(6),式中各变量的含义见表 1.

$$\theta_t{}^{d_{ti}}_{z_t=j} = \frac{n^{d_{ti}}_{z_t=j} + \alpha}{n^{d_{ti}}_{\cdot} + K\alpha} \tag{3}$$

$$\theta_c{}^{d_{ic}}_{z_c=j} = \frac{n^{d_{ic}}_{z_c=j} + \alpha}{n^{d_{ic}}_{\cdot} + L\alpha} \tag{4}$$

$$\delta^{z_t=j}_{w_t} = \frac{n^{z_t=j}_{w_t} + \beta}{n^{z_t=j}_{\cdot} + N\beta} \tag{5}$$

* 本文提到的正文即正常情况下嵌有目录信息的正文.

$$\varphi_{w_c}^{z_c=j}=\frac{n_{w_c}^{z_c=j}+\beta}{n_{\cdot}^{z_c=j}+M\beta} \tag{6}$$

表 1 公式(3~6)各变量含义

$n_{z_t}^{d_{iu}=j}$	文档 i 中的正文信息中分配给主题 z_t 的词汇个数	$n_{z_c}^{d_{ic}=j}$	文档 i 中的目录信息分配给主题 z_c 的词汇个数
$n_{\cdot}^{d_{iu}}$	文档 i 中的正文信息的词汇个数	$n_{\cdot}^{d_{ic}}$	文档 i 中目录信息的词汇个数
$n_{w_t}^{z_t=j}$	词汇 w_t 分配给主题 z_t 的次数	$n_{!}^{z_t=j}$	分配给主题 z_t 的所有词汇个数
$n_{w_c}^{z_c=j}$	词汇 w_c 分配给主题 z_c 的次数	$n_{\cdot}^{z_c=j}$	分配给主题 z_c 的所有词汇个数

3.3 聚类实现

由公式(2)可得文档主题相关度主要由概率 $p(z|d_i)$ 计算得到,但 TC_LDA 建模后得到的主题相关度计算依据包含 2 个因素,即 $p(z_t|d_i)$ 和 $p(z_c|d_i)$. 于是,本文处理方法如下:先根据公式(3)分别处理得到基于图书正文文本的聚类结果集合 R_1 和基于目录信息的聚类结果集合 R_2,再在其基础上计算最终的结果 $R=R_1\cup R_2$.

4 实验

4.1 数据测试集

为了有效验证本文提出的方法的合理性,选取了以下两种语料库:

(1)中文文本分类语料(http://www.nlp.org.cn). 该语料库分为 10 类,属于普通文档. 类别包括环境、计算机、交通、教育、经济、军事、体育、医药、艺术、政治. 我们随机从每个类中选 1000 篇文档进行实验.

(2)图书语料库. 该语料库由我们构建,共包含 150 册数字图书,主要来自于 Internet. 共分为 7 类,分别为管理、健身、经济、军事、旅游、饮食、哲学.

上述两种语料库在实验前都需经过预处理. 首先用汉语词法分析系统 ICTCLAS(http://ictclas.org/)进行分词以及词性标注,再进行停用词的过滤,抽取名词类和动名词类的词汇,并提取各图书的目录信息. 我们在抽取名词和动名词的基础上,过滤掉语料库中出现频数小于 5 的词汇. 同时,聚类效果的评价采取 F1-score[7] 以及准确率(Precision)[7]的测评方法. 其中,F1-score 是准确率(Precision)和查全率(recall)的综合测评指标.

4.2 基于 LDA 模型的聚类算法性能讨论

4.2.1 参数

以中文文本分类语料库为测试数据集,将其效果和目前使用较广泛的聚类算法进行比较,包括 kmeans,FCM(fuzzy c-means)以及 AP(Affinity Propagation)算法[8].

LDA 模型中需设定的参数包括参数 α,β 值、迭代次数以及主题数 K. 由于 α,β 存在经验值[9],因此本文令 $\alpha=50/K,\beta=0.01$(此为经验值,该取值在本语料库上有较好效果). 其中,Gibbs 抽样中迭代的次数又叫 burn-in 间距,目前没有自动选取 burn-in 间距的方法. 针对该参数,本文在随机抽取的 500 篇文档中经多次实验,以 10 为步长,值的选取范围设定为[50,200]. 发现当取值高于 100 时,相对应的聚类结果(准确率)相差较小,差值仅在[-0.03,0.03]之间浮动. 因此,为同时考虑时间消耗和聚类效果两种因素,将 burn-in 间距设为 100.

主题数的选取,会直接影响模型的优劣[10],从而直接影响聚类效果. 文献[11]已证明当建模后主题之间平均相似度最小时,主题模型达到最优. 本文根据该理论,设计了选取最优主题数的方法:(1)首先,根据公式(7)和公式(8)采用 KL 距离计算主题之间的两两相似度,主题数的值选取在范围[100,1000]之内.

$$KL(z_i\parallel z_j)=KL(\delta^{z=i}\parallel\delta^{z=j})=\sum_{k=1}^{N}\delta_{w_k}^{z=i}\log_2\frac{\delta_{w_k}^{z=i}}{\delta_{w_k}^{z=j}} \tag{7}$$

$$dist(z_i,z_j)=\frac{1}{2}[KL(z_i\parallel z_j)+KL(z_j\parallel z_i)] \tag{8}$$

(2)设主题 z_i,z_j 的距离值为 $dist(z_i,z_j)$,主题数为 K,则主题间的平均相似度 avg_dist 定义如下:

$$avg\_dist=\frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}dist(z_i,z_j)}{K\times(K-1)/2} \tag{9}$$

(3)当值 avg_dist 达到最大值时,选取相应的主题数.

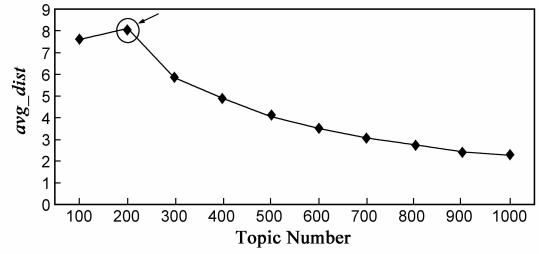

图2 中文文本分类语料库主题数选择实验结果

由上图可得,当最佳主题数设为 200 时,avg_dist 值达到最大,主题之间的相似度最小,则模型达到最优.

4.2.2 各算法聚类效果比较

针对 3 种传统聚类方法,一致采用 TF-IDF 权重机制对文档进行表达. 同时为了比较降维在文本聚类中的效果,我们采用基于流行学习的 ISOMAP 降维方法处理. 每种方法的 F1_score 和准确率(precision)分别记录在表 2 和图 3.

表 2 中,"Kmeans + *"表示在直接使用 Kmeans 方法进行聚类前,需将文档特征向量经过 ISOMAP 算法进行降维处理. 图 3 显示了不同方法在不同数量测试集下

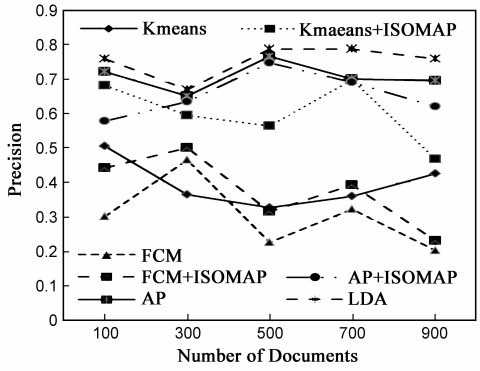

图3　不同方法在不同数量测试集下的准确率(precision)

的准确率.由表 2 和图 3 的数据表明,经过降维处理的 Kmeans 和 FCM 算法的聚类效果优于无降维处理的结果.但 AP 算法的结果优于 AP + ISOMAP,这是因为 AP 算法针对稀疏距离矩阵的聚类效果要好于大规模致密的距离矩阵[8].由所有的方法的比较可得,基于 LDA 的聚类算法效果最优.

表 2　不同聚类方法的 F1_score 值

	Kmeans + *	FCM + *	AP + *	Kmeans	FCM	AP	LDA
F1_score	0.202	0.167	0.185	0.185	0.145	0.215	0.737

图 3 的结果进一步表明 LDA 聚类算法的稳定性,即在不同文档集数目的情况下,LDA 聚类算法始终保持最优.综合表 2 和图 3 的数据可得,基于主题模型的 LDA 方法其优势体现在:提出以主题分析角度进行聚类分析比仅仅基于 BOW(Bag of Words)模型的传统聚类方法效果好,这刚好满足了聚类的本质因素:即同类中的文档共享同一主题.

4.3　基于 TC_LDA 模型的聚类算法讨论

该模型的参数 α,β 同样设定为经验值.迭代次数的选取同 LDA 模型的选取方法相同,并根据实验结果将该值设定为 100.由于 TC_LDA 模型的对象包含正文和目录两部分信息,其主题的个数需不同.我们采取分开确定主题数的方法,分别按照章节 4.2.1 设计的步骤进行.由实验结果进行选取,K 和 L 的最佳取值需设定为 100 和 20.

在表 3 中,我们将 TC_LDA 方法和 LDA 方法,PLSA 以及 AP 算法进行比较.除此之外,TC_LDA 方法的鲁棒性讨论在图 4 中给出.该处实验针对 LDA 方法将主题数设定为 100.针对 PLSA 方法,通常将主题数设定为实际的类别数.在 PLSA 方法中,一旦模型参数学习后,每本书的类别判定通过最大化后验概率实现:

表 3　针对图书语料库四种聚类算法效果比较

算法	AP	PLSA	LDA	TC_LDA
F1_score	0.27	0.68	0.71	0.84

$$cluster(d_i) = \arg\max_{j\in[1\cdots K]} p(z = j \mid d_i) \qquad (10)$$

表 3 将基于主题分析的聚类算法和传统聚类方法中效果相对较好的 AP 算法进行对比.由表 3 的数据可得,基于主题模型的三种方法效果都一致优于 AP 算法.比较三种主题模型方法,TC_LDA 方法的效果最优,其优势主要表现在:(1)TC_LDA 相对于 AP 方法有明显的提高(F1_score 值从 0.27 提高到 0.84);(2)同时结合了目录信息主题建模和正文信息主题建模的 TC_LDA 方法比仅仅针对正文信息主题建模的 LDA 和 PLSA 更能体现优越性(F1_score 分别有 0.13 和 0.16 的提升).

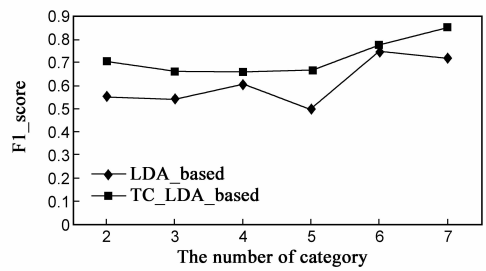

图4　类别数对分类性能的影响

图 4 针对 LDA 和 TC_LDA 分别在不同文档集类别数下的聚类效果进行了剖析.图中横坐标代表不同的类别数,其所含的文档数也线性增长.由图 4 可得,TC_LDA 方法的优势体现在:针对不同数量的测试集,TC_LDA 方法性能一致高于 LDA 方法的聚类效果,从而体现该方法具备一定的鲁棒性.

4.4　时间消耗分析

本文针对每种算法的运行时间进行了分析比较(见表 4).所有算法均运行于同台双核、2GHz 的计算机中.由于每次运行时间会有一定的波动,我们将每种算法运行 10 次,取平均时间进行记录.

表 4　图书语料库中不同算法的时间消耗

算法	LDA	TC_LDA	Kmeans	FCM	AP	PLSA
Time/sec	71.79	78.03	51.14	54.31	44.28	63.83

由上述比较可得,基于主题模型的聚类方法比传统方法需要较多的时间,但差距并不明显.接下来针对 TC_LDA 和 Kmeans 进行详细分析.TC_LDA 中的 Gibbs 采样的时间消耗为 $O(I_L * K * M * \bar{d}_i)$,其中 I_L 为迭代次数,K 为主题数,M 为总的文档数,$\bar{d}_i$ 为每篇文档的平均词汇数.关于 Kmeans,其时间消耗为 $O(I_K * M * C * \bar{C}_w)$,其中 I_K 为迭代次数,C 为类别数,$\bar{C}_w$ 为每个类别中的平均词汇数.各参数的详细比较如下:(1)$I_L > I_K$,因为在 Kmeans 算法中将最大迭代数设定为 100(一般实际运行的迭代数少于该数),而 I_L 设定值为 100.(2)$K > C$,因为 C 被设定为实际的类别数.(3)$\bar{C}_w > \bar{d}_i$,由于 $\bar{C}_w$ 一般由 C 决定一旦聚类数小,一个类别中将包含更多的文档和词汇.

基于上述讨论,无法具体表明哪种算法在时间消耗分析上更具明显优势.同时,尽管本文提出的算法需要提前设定参数的值,但由于在参数估计前设定,并不影响聚类的时间消耗.

5 讨论

本文基于聚类的根本思想,即相同类别的文档共享同一主题,针对普通文档提出基于主题模型(传统LDA模型)的文本聚类方法,实验证明该方法比传统的文档聚类方法具备明显的优势.针对数字图书的特殊语料库,则通过联合数字图书的目录信息和正文信息进行主题建模的方式进行聚类研究,提出了基于 TC_LDA 模型的聚类方法.实验结果表明,基于主题模型的聚类算法比传统方法效果更加优越.在下一步工作中,我们将研究如何在数字图书馆的环境下进一步探讨图书文档间的语义关联.

参考文献

[1] Shehata S, et al. An efficient concept-based mining model for enhancing text clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1360 – 1371.

[2] 刘铭,王晓龙,刘远超.基于语义的高维数据聚类技术[J].电子学报,2009,37(5):925 – 929.
Liu Ming, Wang Xiao-long, Liu Yuan-chao. Clustering technology for high dimensional data based on semantics[J]. Acta Electronica Sinica, 2009, 37(5): 925 – 929. (in Chinese)

[3] Timothy N R, et al. Statistical topic models for multi-label document classification[J]. Machine Learning, 2012, 88(1 – 2): 157 – 208.

[4] Andrzejewski D, Buttler D. Latent topic feedback for information retrieval[A]. Proceedings of 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)[C]. New York: ACM press, 2011. 600 – 608.

[5] Wang X, et al. Topical N-grams: Phrase and topic discovery, with an application to information retrieval[A]. Proc of the 7th IEEE International Conference on Data Mining[C]. Omaha, Nebraska, USA, 2007. 697 – 702.

[6] Heinrich G. Parameter estimation for text analysis[Z/OL]. http://www.arbylon.net/publications/text-est.pdf, 2005.

[7] Ramage D, Heymann P. Clustering the tagged web[A]. Proc of the Second ACM International Conference on Web Search and Data Mining[C]. Barcelona, Spain, 2009. 54 – 63.

[8] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972 – 976.

[9] Blei D M, et al. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993 – 1022.

[10] Newman D, Noh Y, Tally E. Evaluating topic models for digital libraries[A]. Proc of JCDL[C]. Gold Coast, Queensland, Australia, 2010. 215 – 224.

[11] 曹娟,张勇东,李锦涛,唐胜.一种基于密度的自适应最优 LDA 模型选择方法[J].计算机学报,2008,31(10):1780 – 1786.
Cao Juan, Zhang Yong-dong, Li Jin-tao, Tang Sheng. A method of adaptively selecting best LDA model based on density[J]. Chinese Journal of Computer, 2008, 31(10): 1780 – 1787. (in Chinese)

作者简介

王李冬 女,博士研究生,1982 年生于浙江苍南.主要研究方向为图像处理、模式识别、信息检索等.
E-mail: violet_wld@163.com

魏宝刚 男,博士生导师,教授,1960 年生于辽宁沈阳.主要研究领域为人工智 能、图像处理、模式识别等.
E-mail: wbg@zju.edu.cn

袁 杰 男,博士研究生,1981 年生于湖北麻城.主要研究方向为图像处理、模式识别、机器学习、信息检索等.
E-mail: java_mc@163.com