第36卷 第19期 计 算 机 工 程 2010年10月
Vol.36 No.19 Computer Engineering October 2010
·软件技术与数据库· 文章编号：1000—3428(2010)19—0081—03 文献标识码：A 中图分类号：TP301

基于LDA模型的主题词抽取方法

石 晶[1]，李万龙[1,2]
(1. 长春工业大学计算机科学与工程学院，长春 130012；
2. 吉林大学计算机科学与技术学院，长春 130012)

摘 要：以LDA模型表示文本词汇的概率分布，通过香农信息抽取体现主题的关键词。采用背景词汇聚类及主题词联想的方式将主题词扩充到待分析文本之外，尝试挖掘文本的主题内涵。模型拟合基于快速Gibbs抽样算法进行。实验结果表明，快速Gibbs算法的速度约比传统Gibbs算法高5倍，准确率和抽取效率均较高。

关键词：LDA模型；Gibbs抽样；主题词抽取

Topic Words Extraction Method Based on LDA Model

SHI Jing[1], LI Wan-long[1,2]
(1. College of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China;
2. College of Computer Science and Technology, Jilin University, Changchun 130012, China)

【Abstract】Latent Dirichlet Allocation(LDA) is presented to express the distributed probability of words. The topic keywords are extracted according to Shannon information. Words which are not distinctly in the analyzed text can be included to express the topics with the help of word clustering of background and topic words association. The topic meaning is attempted to dig out. Fast Gibbs is used to estimate the parameters. Experiments show that Fast Gibbs is 5 times faster than Gibbs and the precision is satisfactory, which shows the approach is efficient.

【Key words】Latent Dirichlet Allocation(LDA) model; Gibbs sampling; extraction of topic words

1 概述

为了提高主题词提取的效率，本文基于LDA(Latent Dirichlet Allocation)模型[1-3]为语料库及文本建模，利用快速Gibbs抽样进行推理，间接计算模型参数，获取词汇的概率分布。依照香农信息提取片段主题词，通过语料库的词汇聚类产生联想。实验表明以该方法抽取文本的主题词，其结果基本符合人的直觉判断，明显优于其他模型及方法。

2 LDA模型

目前的概率主题模型一般基于同样的思想，即文本是若干主题的随机混合。不同的模型会进一步作不同的统计假设，以不同的方式获取模型参数。

2.1 模型介绍

对于T个主题、D个文本、W个唯一性词汇，文本中的第i个词汇w_i可表示为$P(w_i)=\sum_{j=1}^{T}P(w_i\mid z_i=j)P(z_i=j)$，其中，$z_i$是潜在变量，表明第$i$个词汇记号$w_i$取自该主题；$P(w_i\mid z_i=j)$是词汇$w_i$记号属于主题$j$的概率；$P(z_i=j)$给出主题$j$属于当前文本的概率。假定$T$个主题形成$D$个文本以$W$个唯一性词汇表示，为记号方便，令$\varphi_w^{(z=j)}=P(w\mid z=j)$表示对于主题$j$，$W$个词汇上的多项分布，其中，$w$是$W$个唯一性词汇表中的词汇；令$\psi_{z=j}^{(d)}=P(z=j)$表示$T$个主题上对于文本$d$的多项分布，于是文本$d$中词汇$w$的概率为$P(w\mid d)=\sum_{j=1}^{T}\varphi_w^{(z=j)}\cdot\psi_{z=j}^{(d)}$。

LDA模型在$\psi^{(d)}$上作$Dirichlt(a)$的先验概率假设，使模型易于处理训练语料之外的新文本。为了便于模型参数的推理，本文除了在$\psi^{(d)}$上作对称的$Dirichlt(a)$的先验概率假设外，在$\varphi^{(z)}$上亦作如下对称的$Dirichlt(X)$的先验概率假设[4-5]：

$$w_i\mid z_i,\varphi^{(z_i)}\sim Discrete(\varphi^{(z_i)}),\quad \varphi^{(z_i)}\sim Dirichlet(\chi),$$
$$z_i\mid\psi^{(d_i)}\sim Discrete(\psi^{(d_i)}),\quad \psi^{(d_i)}\sim Dirichlet(\alpha)$$

2.2 快速Gibbs抽样

为了获取词汇的概率分布，本文利用快速Gibbs抽样间接求得φ和ψ的值。对于本文的LDA模型，仅仅需要对主题的词汇分配，也就是对变量Z_i进行抽样。记后验概率为$P(z_{i,d}=j\mid z_{-i,d},w_{i,d},\alpha,\chi)$，计算公式如下(计算细节见文献[6])：

$$P(z_{i,d}=j\mid z_{-i,d},w_{i,d},\alpha,\chi)=\frac{1}{Nom}\cdot\frac{(n_{-i,j}^{(w_i)}+\chi)(n_{-i,j}^{(d_i)}+\alpha)}{n_{-i,j}^{(\cdot)}+W\chi}\tag{1}$$

其中，Nom是标准化因子，$Nom=\sum_{j=1}^{T}\frac{(n_{-i,j}^{(w_i)}+\chi)(n_{-i,j}^{(d_i)}+\alpha)}{n_{-i,j}^{(\cdot)}+W\chi}$；$z_{i,d}=j$表示将词汇记号$w_i$分配给主题$j$；$z_{-i,d}$表示所有$z_{k,d}(k\neq i)$的分配；$n_{-i,j}^{(w_i)}$是分配给主题$j$与$w_i$相同的词汇个数；$n_{-i,j}^{(\cdot)}$是分配给主题$j$的所有词汇个数；$n_{-i,j}^{(d_i)}$是文本$d_i$中分配给主题$j$的词汇个数；所有的词汇个数均不包括这次$z_i=j$的分配。

舍弃词汇记号，以w表示唯一性词，对于每一个单一样

基金项目：长春工业大学博士基金资助项目(2008A02)
作者简介：石 晶(1970－)，女，讲师、博士，主研方向：中文信息处理；李万龙，教授
收稿日期：2010-03-09 **E-mail：**crystal1087@126.com

本，可以按下式估算 φ 和 ψ 的值：

$$\hat{\varphi}_w^{(z=j)} = \frac{n_j^{(w)} + \chi}{n_j^{(\cdot)} + W\chi}, \quad \hat{\psi}_{z=j}^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha} \tag{2}$$

其中，$n_j^{(w)}$ 表示词汇 w 被分配给主题 j 的频数；$n_j^{(\cdot)}$ 表示分配给主题 j 的所有词数；$n_j^{(d)}$ 表示文本 d 中分配给主题 j 的词数；$n_\cdot^{(d)}$ 表示文本 d 所有被分配了主题的词数。

与文献[7]类似，本文定义向量 $\boldsymbol{a}$、$\boldsymbol{b}$、$\boldsymbol{c}$ 为：

$$\boldsymbol{a} = [n_{-i,1}^{(d)} + \alpha, n_{-i,2}^{(d)} + \alpha, \cdots, n_{-i,k}^{(d)} + \alpha],\ \boldsymbol{b} = [n_{i,1}^{(w_i)} + X, n_{i,2}^{(w_i)} + X, \cdots, n_{i,k}^{(w_i)} + X]$$

$$\boldsymbol{c} = [1/(n_{-i,1}^{(\cdot)} + W\chi), 1/(n_{-i,2}^{(\cdot)} + W\chi), \cdots, 1/(n_{-i,k}^{(\cdot)} + W\chi)],\ p_k = \boldsymbol{a}_k\boldsymbol{b}_k\boldsymbol{c}_k,\ Nom = \sum_k p_k$$

根据 Holder 不等式得到 Nom 的初始限制值：$Nom_0 = \|\boldsymbol{a}\|_p \|\boldsymbol{b}\|_q \|\boldsymbol{c}\|_r \geq Nom$，其中，$1/p + 1/q + 1/r = 1$，第 l 步的限制值为 $Nom_l = \sum_{i=1}^{l}(\boldsymbol{a}_i\boldsymbol{b}_i\boldsymbol{c}_i) + \|\boldsymbol{a}_{l+1:K}\|_p \|\boldsymbol{b}_{l+1:K}\|_q \|\boldsymbol{c}_{l+1:K}\|_r \geq Nom$。

快速 Gibbs 抽样算法步骤如下(传统 Gibbs 抽样算法参见文献[6])：

对于每个词汇记号 $w_i(1 \leq i \leq N$，N 为所有词汇记号总数)：

(1) μ 被初始化为 0~1 之间的随机数，同时令 $inip_0$=0；

(2)对于每一个主题 $j(1 \leq j \leq T)$：1)$inip_j$=$inip_j$－1+p_j；2)根据式(6)计算 Nom_j；3)如果 $\mu \times Nom_j \leq initp_j$。

①若 j=1 或者 $\mu \times Nom_j > initp_{j-1}$，将 j 分配给 w_j，否则，对于每个主题 $k(1 \leq k \leq T)$，$\mu = \dfrac{(\mu \times Nom_{j-1} - initp_{j-1}) \times Nom_j}{Nom_{j-1} - Nom_j}$。

②当 $\mu \leq initp_k$，将 k 分配给 w_j。

3 词汇聚类

仅仅依赖所在文本的内部信息确定主题词，错误较多，如果能够借助背景库使主题词产生联想，必然有助于准确率的提高，为此需要利用丰富的背景库知识聚类词汇。本文以 1998 年《人民日报》手工标注的语料为背景库，以知网词典中的每一个词作为种子词，选择与之最相关的 n 个词形成一个聚类。对于每一个词汇 w，按下式计算该词汇对于种子词 s 的 δSC 值，根据 MDL 原则[4]，δSC 值越大，w 与 s 的相关性越强。

$$\delta SC = H\left(\frac{m^+}{m}\right) - \frac{m_s}{m}H\left(\frac{m_s^+}{m_s}\right) - \frac{m_{\neg s}}{m}H\left(\frac{m_{\neg s}^+}{m_{\neg s}}\right) - \frac{1}{2m}\ln\left(\frac{m_s m_{\neg s}^{\pi}}{2m}\right) \tag{3}$$

其中，$H(z) = -z\ln(z) - (1-z)\ln(1-z), 0 < z < 1$，当 $z = 0$ 或 $z = 1$ 时，$H(z) = 0$；m^+ 表示出现 w 的文本数；m_s 表示出现 s 的文本数；m_s^+ 表示 w、s 共现的文本数；$m_{\neg s}$ 表示不出现 s 的文本数；$m_{\neg s}^+$ 表示出现 w 但不出现 s 的文本数；m 表示总的文本数。

4 主题分析

4.1 提取方法

根据公式 $P(w \mid t) = \sum_{j=1}^{T} \varphi_w^{(z=j)} \cdot \psi_{z=j}^{(t)}$ 求得文本 t 的词汇概率分布。利用该概率分布，定义词汇 w 在文本 t 中的香农信息[8]：

$$I(w) = -N(w)\ln P(w \mid t) \tag{4}$$

其中，$N(w)$是 w 在文本 t 中的出现频数。香农信息值越大，说明其在该片段中的价值越大，于是选择香农信息较大的一个词汇形成主题词串，代表该片段的主题。由于式(4)的词汇概率分布不仅蕴含语料库学习的知识，而且反映被提取片段的信息，因此有助于提高主题词提取的准确率。

4.2 主题词联想

在词汇聚类表中选择种子词是主题词的聚类，并使主题词根据背景词汇聚类产生联想。联想包括归一、合并、替换 3 个过程。

归一：令 2 个聚类分别为 $(s: w_1, w_2, \cdots, w_n)$、$(s': w_1', w_2', \cdots, w_m')$，其中，$s$、$s'$是种子词；$w_i(1 \leq i \leq n)$、$w_j'(1 \leq j \leq m)$是 2 个聚类中的非种子词，若 $s = w_j'$，$s' = w_i$，$(1 \leq i \leq n, 1 \leq j \leq m)$，且至少 s、s'之一为主题词，则将主题词扩充为 $s - s'$，s、s'被称为主题词元素，处理后的主题词形成归一主题词表。

合并：若 2 个主题词含有公共元素，则将这 2 个主题词合并为一个主题词。即对于主题词 $A - s - B$、$A' - s - B'$，合并两者为 $A - s - B - A' - B'$，其中，A, B, A', B' 可能由多个主题词元素构成。该过程遍历归一主题词表的所有主题词，直至没有重复的主题词元素，处理后的主题词表称为合并主题词表。

替换：若合并主题词表有主题词 $s - s'$，而原主题词表中有 s 或 s'，将其替换为 $s - s'$，循环此过程直到合并主题词表中的所有主题词均得以替换，原主题词表中的其他主题词保持不变，形成替换主题词表。

最后删除替换主题词表中重复的主题词，形成新的主题词表，该主题词表即为经过背景词汇聚类联想后的主题词表。

5 实验设计及结果对比

本文所有实验以 1998 年《人民日报》手工标注的语料库为背景库及建模对象(共 3 157 个文本)，并以知网词典(去除其中的虚词、形容词、副词等意义不大的词，再删掉语料库出现频数小于 5 的词，剩余 18 049 个词汇)作为选择词汇的词典。

5.1 词汇聚类

以词典中的每一个词汇作为种子词 s，当 $P(w \mid s) > P(w)$ 时，取 7 个 $\delta SC > \gamma, \gamma = 0.005$ 的词汇(按 δSC 值从大到小的顺序)和 3 个 $rel(w,s) > \gamma', \gamma' = 0.002\ 5$ 的词汇(按 $rel(w,s)$值从大到小的顺序)，构成同一个聚类。舍弃独词(只包括种子词)聚类，形成词汇聚类表。共有 6 502 个聚类出现。

5.2 快速 Gibbs 与 Gibbs 的速度对比

Gibbs 抽样可以用于拟合 LDA 模型，但速度较慢。本文采用快速 Gibbs 抽样算法，速度提高近 5 倍，详见图 1。

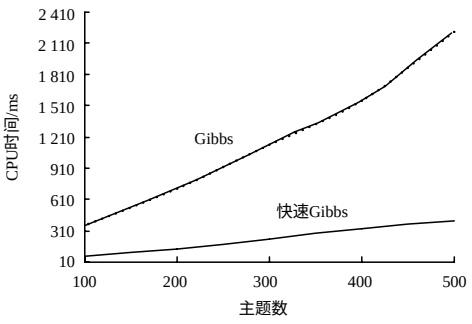

图 1 快速 Gibbs 与 Gibbs 的速度比较

5.3 主题提取的测试

5.3.1 测试语料

本测试采用的是文本分类语料库，共包括环境、经济、

艺术、教育、体育、计算机、医学、政治、交通、军事等十大类。测试语料库中的文本没有分词，所以首先利用北大的分词系统 ICTCLAS 对其进行处理，然后凭直觉给予每个类一定数量(最多 5 个)的标识词，见表 1。

表 1　类及其标识词

类	标识词
环境	环境 动物 土壤 植被
经济	经济 金融 财政 商品 贸易
军事	战 军 炸弹 航空 装备
计算机	电脑 网 芯片 数据 程序
交通	交通 车 乘客 路 港口
教育	教育 思想 校 学
体育	赛 训练
艺术	文艺 艺术 拍摄 出版 剧院
医药	病 伤口 药 饮食
政治	政治 会 访问 联合国 和平

5.3.2　度量标准

若从某类文本提取的主题词包含该类的标识词，即认为提取结果正确。准确率定义为：$precision=\frac{n_{\text{correct}}}{n_{\text{total}}}$，其中，$n_{\text{conect}}$ 指正确提取主题词的文本数；n_{total} 指测试文本的总数。

5.3.3　主题提取

Gibbs 抽样的主题数目 T=80，超参数 a=50/T, X=0.01，取 10 个不同的初始值运行算法，每个初始值迭代 1 000 次，然后每隔 100 次取一次样本(取值的具体方法见文献[2])，共取 10 次样本。加入训练语料的测试文本被初始化，继续迭代 10 次，开始计算结果。每个文本的测试结果取 100 个样本的平均值，测试集的实验结果取所有文本测试结果的平均值，以此计算词汇的分布概率。主题提取时以类为单位进行测试，每个类取约 100 个主题片段，其测试集合如表 2 所示。

表 2　测试集及所包含的片段数目

类	主题词
环境	101
经济	113
军事	97
计算机	95
交通	102
艺术	114
医药	106
政治	105
教育	100
体育	105

将本文方法与 TF-IDF[9]及 Z-SCORE[9]方法进行对比。TF-IDF 的计算方法为：

$$weight_w(s)=\frac{tf_w(s)\times \text{lb}(\frac{N}{n_w})}{\sqrt{\sum_{i=1}^{n}(tf_w(s))^2\times \text{lb}(\frac{N}{n_w})}} \tag{5}$$

其中，$tf_w(s)$ 表示词汇 w 在测试片段 s 中的出现频数；N 为背景语料中所有的片段数目；n_w 是背景语料含有 w 的片段数目。Z-SCORE 的计算方法为：

$$weight_w(s)=\frac{tf_w(s)-\frac{\sum_{i=1}^{N}f_i(w)}{N}}{\sqrt{\sum_{i=1}^{N}\left(f_i(w)-\frac{\sum_{i=1}^{N}f_i(w)}{N}\right)^2\times\frac{f_i(w)}{N}}} \tag{6}$$

其中，$f_i(w)$ 是词汇 w 在背景语料第 i 个片段中的出现频数。

当提取主题词的数量为 5 时，实验结果如表 3 所示。从表中可见，本文方法的结果远远好于其他 2 种方法，主要原因在于充分利用了背景语料库的知识，使主题词产生联想，以此挖掘出隐藏于文本之中的内涵。

表 3　主题词数为 5 的提取结果　(%)

Category	Shannon	TF-IDF	Z-SCORE
环境	79.82	47.17	9.44
经济	92.16	41.11	29.68
军事	70.00	50.52	10.17
计算机	70.91	56.37	24.33
交通	100.00	79.69	33.12
教育	98.04	80.00	21.98
体育	95.31	54.69	76.65
艺术	98.15	54.72	6.64
医药	90.00	61.67	44.83
政治	94.55	62.13	30.10

6　结束语

本文利用 LDA 为语料库及文本建模，通过背景知识解析文本的主题。LDA 是完全的生成模型，从理论上讲，具有其他模型无可比拟的建模优点。为了提高主题词提取的效率，本文以词汇聚类的方式使主题词产生联想，将主题词扩充到待分析文本之外，尝试挖掘隐藏于字词表面之后的文本内涵，并利用快速 Gibbs 抽样算法对模型进行拟合。实验结果表明，本文方法有很好的分析表现，能够为下一步文本推理的研究提供坚实的基础。

参考文献

[1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] Li Wenbo, Sun Le, Zhang Dakun. Text Classification Based on Labeled-LDA Model[J]. Chinese Journal of Computers, 2008, 31(4): 620-627.

[3] Caol J, Li Jintao, Zhang Yongdong, et al. LDA-based Retrieval Framework for Semantic News Video Retrieval[C]//Proc. of Conf. on Semantic Computing. Irvine, California, USA: IEEE Press, 2007.

[4] Steyvers M, Griffiths T. Probabilistic Topic Models[M]//Landauer T, McNamara D, Dennis S, et al. Latent Semantic Analysis: A Road to Meaning. [S. l.]: MIT Press, 2006.

[5] Griffiths T, Steyvers M. Finding Scientific Topics[J]. Proceedings of the National Academy of Sciences, 2004, 101(Suppl. 1): 5228-5235.

[6] Shi Jing, Hu Ming, Shi Xin, et al. Text Segmentation Based on Model LDA[J]. Chinese Journal of Computers, 2008, 31(10): 1865-1873.

[7] Nevada L V. Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2008: 569-577.

[8] Li Hang, Yamanishi K. Topic Analysis Using a Finite Mixture Model[J]. Information Processing & Management, 2003, 39(4): 521-541.

[9] Liu Ying, Ciliax B J, Borges K, et al. Comparison of Two Schemes for Automatic Keyword Extraction from MEDLINE for Functional Gene Clustering[C]//Proc. of IEEE Computational Systems Bioinformatics Conference. Stanford, California, USA: IEEE Press, 2004: 394-404.

编辑　张正兴