基于 LDA 模型的主题分析

石 晶 [1] 范 猛 [2] 李万龙 [1,3]

摘 要 在文本分割的基础上, 确定片段主题, 进而总结全文的中心主题, 使文本的主题脉络呈现出来, 主题以词串的形式表示. 为了分析准确, 利用 LDA (Latent dirichlet allocation) 为语料库及文本建模, 以 Clarity 度量块间相似性, 并通过局部最小值识别片段边界. 依据词汇的香农信息提取片段主题词, 采取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, 尝试挖掘隐藏于字词表面之下的文本内涵. 实验表明, 文本分析的结果明显好于其他方法, 可以为下一步文本推理的工作提供有价值的预处理.

关键词 主题分析, LDA 模型, 文本分割, Gibbs 抽样
中图分类号 TP301

Topic Analysis Based on LDA Model

SHI Jing[1] FAN Meng[2] LI Wan-Long[1,3]

Abstract Topic spotting of segments is performed based on text segmentation and the main topic of the whole text is then generalized. Topics are represented by means of word clusters. LDA (Latent dirichlet allocation) is used to model corpora and text. Clarity is taken as a metric for similarity of blocks and segmentation points are identified by local minimum. The topic words of segments 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 signification behind the words are attempted to be digged out. Experiments tell that the result of analyzing is far better than those of other methods. Valuable pre-processing is provided for text reasoning.

Key words Topic analysis, latent dirichlet allocation (LDA) model, text segmentation, Gibbs sampling

文本的主题分析旨在确定一个文本的主题结构, 即识别所讨论的主题, 界定主题的外延, 跟踪主题的转换, 觉察主题间的关系等, 分析结果对于信息提取、文摘自动生成、文本分类等领域都有极为重要的价值. 主题分析的程度随着应用对象的不同有所区别, 浅层次的分析仅仅确定主题边界 (文本分割)[1−2], 或者进而指明不同片段间的关系 (是否讨论同一主题)[3]; 比较复杂的分析能够在识别边界的基础上讨论主题的内容[4]. 作为文本推理的预处理, 本文研究如何将边界计算及主题表示集中在 LDA (Latent dirichlet allocation) 模型的框架下统一实现.

欲利用统计的方法分析文本, 首先必须选择合适的模型. 文献 [4] 以不附加任何统计假设的有限混合模型 (Finite mixture model) 代表文本中的词汇分布, 直接利用 EM(Expectation maximization) 对其进行训练, 导致的问题

收稿日期 2008-07-16 收修改稿日期 2009-03-25
Received July 16, 2008; in revised form March 25, 2009
长春工业大学博士基金 (2008A02) 资助
Supported by Changchun Technology University Doctoral Program (2008A02)
1. 长春工业大学计算机科学与工程学院 长春 130012 2. 长春工业大学科研处 长春 130012 3. 吉林大学计算机科学与技术学院 长春 130012
1. College of Computer Science and Engineering, Changchun University of Technology, Changchun 130012 2. Department of Science and Research Administration, Changchun University of Technology, Changchun 130012 3. College of Computer Science and Technology, Jilin University, Changchun 130012
DOI: 10.3724/SP.J.1004.2009.01586

是出现局部极大值, 且收敛速度过慢. 文中假定不同的片段由不同的模型产生, 每个模型单独训练, 意味着分割时仅仅依靠本文档的信息, 并不吸收语料库学习的知识, 错误率必然较高. PLSA (Probabilistic latent semantic analysis)[5] 是另一可选模型, 但模型中的文档概率值与特定文档相关, 因此缺乏处理新文档的自然方法. 同时待估参数的数量随着文档数量的增多线性增长, 说明模型易于过度拟合. 与 PLSA 模型相比, LDA (Latent dirichlet allocation)[6] 称得上是完全的生成模型. 由于该模型将主题混合权重视为 k 维参数的潜在随机变量, 而非与训练数据直接联系的个体参数集合, 推理上采用 Laplace 近似, 变分近似[6], MCMC (Markov chain Monte Carlo)[7] 以及期望–扩散 (Expectation-propagation)[8] 等方法获取待估参数值, 所以克服了上述不足.

本文基于 LDA 模型为语料库及文本建模, 利用 MCMC 中的 Gibbs 抽样进行推理, 间接计算模型参数, 获取词汇的概率分布. 然后: 1) 根据 Clarity 度量 (基于词汇的概率分布值) 求得句间相似性, 以局部最小值的方式识别片段边界; 2) 依照香农信息 (利用词汇的概率分布值计算) 提取片段主题词, 并通过语料库的词汇聚类产生联想; 3) 从联想后的片段主题词中提取全文中心主题词. 由于充分利用了语料库的词汇聚类使主题词产生联想, 从而大幅度提高了主题词提取的准确率. 实验结果表明以该方法分析文本的主题脉络, 其结果基本符合人的直觉判断, 且明显优于其他模型及方法.

本文的结构安排如下, 第 1 节介绍 LDA 模型. 第 2 节和第 3 节介绍基于 LDA 模型的主题分析方法, 其中第 2 节给出文本分割的策略; 第 3 节详述在文本分割的基础上实现主题提取的方法以及如何通过词汇聚类提高提取的准确率. 第 4 节给出测试手段及实验结果, 并就实验结果进行讨论. 第 5 节对比、分析相关研究及工作. 最后总结全文.

1 LDA 模型

目前的概率主题模型一般基于同样的思想–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.

1.1 模型介绍

一个文本通常需要讨论若干主题, 而文本中的特定词汇体现出所讨论的特定主题. 在统计自然语言处理中, 为文本主题建模的方法是视主题为词汇的概率分布, 文本为这些主题的随机混合. 假设有 T 个主题, 则所给文本中的第 i 个词汇 w_i 可以表示如下:

$$P(w_i)=\sum_{j=1}^{T}P(w_i|z_i=j)P(z_i=j) \tag{1}$$

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, $P(w_i|z_i=j)$ 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 $\phi_w^{(z=j)}=P(z=j)$ 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 $\psi_{z=j}^{(d)}=P(z=j)$ 表示对于文本 d, T 个主题上的多项分布, 于是文本 d 中词汇 w 的概率为:

$$P(w|d)=\sum_{j=1}^{T}\phi_w^{(z=j)}\cdot\psi_{z=j}^{(d)} \tag{2}$$

LDA 模型[6] 在 $\psi^{(d)}$ 上作 $Dirichlet(\alpha)$ 的先验概率假设, 使得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 $Dirichlet(\alpha)$ 的先验概率假设外, 在 $\phi^{(z)}$ 上亦作对称的 $Dirichlet(\chi)$ 的先验概率假设[9], 如下:

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

这里的 χ 可以理解为, 在见到语料库的任何词汇之前, 从主题抽样获得的词汇出现频数, 而 α 可以理解为, 在见到任何文档文字之前, 主题被抽样的频数. 尽管 χ 和 α 的具体取值会影响到主题及词汇被利用的程度, 但不同的主题被利用的方式几乎没有变化, 不同的词汇被利用的方式也基本相同, 因此可以假定对称的 Dirichlet 分布, 即所有的 χ 取相同的值, 所有的 α 取相同的值.

1.2 Gibbs 抽样

为了获取词汇的概率分布, 本文没有将 ϕ 和 ψ 作为参数直接计算, 而是考虑词汇对于主题的后验概率 $P(w|z)$, 利用 Gibbs 抽样间接求得 ϕ 和 ψ 的值. MCMC 是一套从复杂的概率分布抽取样本值的近似迭代方法, Gibbs 抽样作为 MCMC 的一种简单实现形式, 其目的是构造收敛于某目标概率分布的 Markov 链, 并从链中抽取被认为接近该概率分布值的样本. 于是目标概率分布函数的给出便成为使用 Gibbs 抽样的关键. 对于本文的 LDA 模型, 仅仅需要对主题的词汇分配, 也就是变量 z_i 进行抽样. 记后验概率为 $P(z_i=j|z_{-i},w_i)$, 计算公式如下:

$$P(z_i=j|z_{-i},w_i)=\frac{\dfrac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+W\chi}\cdot\dfrac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha}}{\sum\limits_{j=1}^{T}\dfrac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+W\chi}\cdot\dfrac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha}} \tag{3}$$

其中, $z_i=j$ 表示将词汇记号 w_i 分配给主题 j, 这里 w_i 被称为词汇记号是因为其不仅代表词汇 w, 而且与该词所在的文本位置相关, z_{-i} 表示所有 $z_k(k\neq i)$ 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; $n_{-i,\cdot}^{(d_i)}$ 是 d_i 中所有被分配了主题的词汇个数; 所有的词汇个数均不包括这次 $z_i=j$ 的分配.

Gibbs 抽样算法详述如下 (具体理论描述详见文献 [10]):

1) z_i 被初始化为 1 到 T 之间的某个随机整数. i 从 1 循环到 N, N 是语料库中所有出现于文本中的词汇记号个数. 此为 Markov 链的初始状态.

2) i 从 1 循环到 N, 根据式 (3) 将词汇分配给主题, 获取 Markov 链的下一个状态.

3) 迭代第 2) 步足够次数以后, 认为 Markov 链接近目标分布, 遂取 z_i (i 从 1 循环到 N) 的当前值作为样本记录下来. 为了保证自相关较小, 每迭代一定次数, 记录其他的样本. 舍弃词汇记号, 以 w 表示唯一性词, 对于每一个单一样本, 可以按下式估算 ϕ 和 ψ 的值:

$$\tilde{\phi}_w^{(z=j)}=\frac{n_j^{(w)}+\chi}{n_j^{(\cdot)}+W\chi},\quad \tilde{\psi}_{z=j}^{(d)}=\frac{n_j^{(d)}+\alpha}{n_{\cdot}^{(d)}+T\alpha} \tag{4}$$

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 中所有被分配了主题的词数.

Gibbs 抽样算法从初始值开始运行, 迭代足够次 b 后认为样本接近目标概率分布, 然后每隔一定次数 c 抽取样本, b 称为 Burn-in 间距, c 称为 Thinning 间距. b 和 c 的取值比较难于确定, 一般与特定的语料库相关. 如果所构造 Markov 链的相邻状态间关联较小, b, c 以较小的值可满足需要, 但如果相邻状态间的关联较大, 就必须增大 b, c 的取值, 方可降低自相关.

2 文本分割

2.1 分割策略

待分割文本是语料库训练时没有处理过的新文本, 如果对于每一个未知文本, 都将其加入语料库后重新训练, 则异常浪费时间, 也没有必要, 本文的做法是只对新加入的词汇记号运行 Gibbs 抽样算法, 且只迭代较少的次数. 预处理的基本块采用汉语的整句 s, 分割的大致步骤如下:

步骤 1. 对于语料库文本的词汇记号运行 Gibbs 抽样算法, 迭代足够次;

步骤 2. 以整句 s 作为式 (3) 中的文本 d, 遍历待分割文本的所有词汇记号, 运行 Gibbs 抽样算法, 迭代少数几次;

步骤 3. 按照式 (4) 分别计算 ϕ 和 ψ 的值;

步骤 4. 根据公式 $P(w|s)=\sum_{j=1}^{T}\phi_w^{(z=j)}\cdot\psi_{z=j}^{(s)}$ 求取待分割文本词汇的概率分布 $P(w|s)$;

步骤 5. 基于 $P(w|s)$, 利用 Clarity 度量计算句间的相似值 Sim;

步骤 6. 结合局部最小值的边界估计策略, 通过句间相似值 Sim 识别片段边界.

2.2 Clarity 度量

$$\begin{aligned}Sim_{Clr}=&-KL(P(w|s_1)||P(w|s_2))+KL(P(w|s_1)||GC)-\\&KL(P(w|s_2)||P(w|s_1))+KL(P(w|s_2)||GC)\end{aligned}\tag{5}$$

其中, GC 代表词汇 w 在训练语料库的出现频率, 即 $f(w)$, $KL(\cdot||\cdot)$ 被称为相对熵:

$$KL(P(w|s_1)||P(w|s_2))=\sum_{w\in W}P(w|s_1)\log_2\frac{P(w|s_1)}{P(w|s_2)}\tag{6}$$

2.3 局部最小值法[11]

假设待分割文本有 n 个整句, 则相邻句间的相似值表为 $SimTable=\{Sim_1,Sim_2,\cdots,Sim_i,\cdots,Sim_{n-1}\}$, 其中 $Sim_i=Sim(s_i,s_{i+1}),1\le i\le n-1$. 在表中选择局部最小值 $Sim_{\min}(s_1,s_2)$; 从每一个局部最小值出发向左、向右分别寻找距离最近的较大值 $Sim_{\max l}$ 以及 $Sim_{\max r}$, 利用公式 $d_{rel}(s_1,s_2)=\dfrac{Sim_{\max l}+Sim_{\max r}}{2Sim_{\min}(s_1,s_2)}-1$ 计算相对深度; 令 η 为一常数, 若相对深度 $d_{rel}(s_1,s_2)>\eta$, 则 s_1, s_2 分属于不同的片段.

3 主题提取

3.1 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 (Minimum description length) 原则[4], δSC 值越大, 说明 w 与 s 的相关性越大.

$$\begin{aligned}\delta SC=&H\left(\frac{m^+}{m}\right)-\frac{m_s}{m}H\left(\frac{m_s^+}{m_s}\right)-\\&\frac{m_{\neg s}}{m}H\left(\frac{m_{\neg s}^+}{m_{\neg s}}\right)-\frac{1}{2m}\ln\left(\frac{m_s m_{\neg s}\pi}{2m}\right)\end{aligned}\tag{7}$$

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.

本文采用的聚类方法除了考虑种子词与其他词在不同文本的共现, 还进一步考虑二者在同一文本的共现强弱, 以 $rel(w,s)=freq(w,s)/k$ 计算, $freq(w,s)$ 表示在 s 出现的文本中, w 出现的频数, k 是语料库文本的平均词数. $rel(w,s)$ 直接体现出某一文本中 w 与 s 的相关程度. 随着 δSC 的值逐渐减小, 与种子词相关的词越来越难于选出, 或者选择的词与种子词联系得越来越不紧密, 此时以某一文本中与种子词频繁共现的词汇予以补充或替换, 从而使聚类结果更令人满意. 实验表明该方法确实比单独使用 MDL 的方法能够更好地吻合人的直觉, 也有利于主题词的联想.

3.2 片段主题提取

3.2.1 提取方法

文本被分割为若干片段 $t_1,t_2,\cdots,t_k$ 后, 将片段 t_k $(1\le k\le h)$ 作为新的文本加入到语料库中, 依次在每个片段上运行 Gibbs 抽样算法, 根据公式 $P(w|t_k)=\sum_{j=1}^{T}\phi_w^{(z=j)}\cdot\psi_{z=j}^{(t_k)}$ 求得片段 t_k $(1\le k\le h)$ 的词汇概率分布. 利用该概率分布定义词汇 w 在片段 t_k 中的香农信息[4]:

$$I(w)=-N(w)\ln P(w|t_k),\quad 1\le k\le h\tag{8}$$

其中, $N(w)$ 是 w 在片段 t_k 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 l 个词汇形成主题词串, 代表该片段的主题. 由于式 (8) 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, 所以有助于提高主题词提取的准确率.

3.2.2 主题词联想

利用词汇聚类可以使主题词产生联想, 从而提高主题分析的准确性. 具体方法是: 在词汇聚类表中选择种子词是主题词的聚类, 通过归一、合并、替换 3 个过程实现联想.

归一: 令两个聚类分别为 $(s:w_1,w_2,\cdots,w_n)$, $(s':w_1',w_2',\cdots,w_m')$, 其中 s,s' 是种子词, w_i $(1\le i\le n)$, w_i' $(1\le j\le m)$ 是两个聚类中的非种子词, 若 $s=w_i'$, $s'=w_i$, $(1\le i\le n,\ 1\le j\le m)$ 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. 例如两个聚类为: (学生: 学校, 教师, 老师, 中学, 同学, 教学, 教育) 和 (教师: 学生, 教学, 学校, 中小学, 教育, 办学, 中学), 并且 "学生" 为原主题词, 则将主题词扩充为 "学生 − 教师".

合并: 若两个主题词含有公共元素, 则将这两个主题词合并为一个主题词. 即对于主题词 $A-s-B$, $A'-s-B'$,

合并二者为 $A-s-B-A'-B'$, 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, 处理后的主题词表被称为合并主题词表. 例如两个主题词为 "货币 – 金融 – 贸易" 和 "出口 – 贸易 – 危机", 则将其合并为 "货币 – 金融 – 贸易 – 出口 – 危机".

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. 例如原主题词包括 "学生", 而合并主题词表中有 "学生 – 教师", 则以 "学生 – 教师" 替换 "学生".

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.

3.3　中心主题获取

根据联想后的片段主题词计算全文的中心主题词. 假设文本有 n 个片段, 其中 s 作为不同片段的主题词元素出现 m 次, 则 $P(s)=m/n$, 取 $P(s)>\mu$ 的主题词元素作为中心主题词, μ 为一小于 1 的常数.

4　实验设计及结果对比

本文所有实验以 1997 年和 1998 年人民日报手工标注的语料库以及文本分类语料库为背景库及建模对象 (共 12 980 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 10 的词, 剩余 9 768 个词汇) 作为选择词汇的词典. 除知网外, 抽取关键词实验还用到汉语词法分析系统 ICTCLAS. 所有这些实验中用到的资源均可网上下载并限研究使用.

为了有效利用 Gibbs 抽样算法, 先通过实验确定主题数目 T 的最佳值, 以及 Burn-in 间距 (b) 和 Thinning 间距 (c) 的取值. 对于主题结构的测试, 按文本分割及主题提取两个方面分别单独进行.

4.1　词汇聚类

以词典中的每一个词汇作为种子词 s, 当 $P(w|s)>P(w)$ 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 $rel(w,s)>r'$, $r'=0.00025$ 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.

4.2　主题数目的确定

针对同样的语料库及同样的词典 ($W=9\,768$, $D=12\,980$, $N=1\,032\,365$, W 为词汇数目, D 为文本数目, N 为词汇记号数目, 也就是每次抽样依据式 (3) 对 z 赋值的次数), 可变量包括超参数 α,χ 以及主题数目 T. 本实验目的在于了解主题数目对于 Gibbs 抽样算法的影响, 为此先确定 α, χ 的值, 然后为 T 选择合适的值. 这实际上是一个模型选择的问题, 本文采用贝叶斯统计中的标准方法予以解决. 令 $\alpha=50/t$, $\chi=0.01$ (此为经验值, 多次实验表明, 这种取值在本实验的语料库上有较好表现), T 取不同的值分别运行 Gibbs 抽样算法, 检测 $\ln P(w|T)$ 值的变化.

由本文建模的模型可知, α, χ 是多项分布 ψ 和 ϕ 上的 Dirichlet 先验概率假设, 其自然共轭的特点说明通过对 ψ 和 ϕ 积分可以求取联合概率 $P(w,z)$ 的值. $P(w,z)=P(w|z)P(z)$, 并且 ϕ 和 ψ 分别单独出现于第 1 项和第 2 项, 对 ϕ 积分获第 1 项值如下:

$$P(w|z)=\left(\frac{\Gamma(W\chi)}{\Gamma(\chi)^W}\right)^{\mathrm{T}}\prod_{i=1}^{T}\frac{\prod_w\Gamma(n_j^{(w)}+\chi)}{\Gamma(n_j^{(\cdot)}+W\chi)} \tag{9}$$

其中, $\Gamma(\cdot)$ 是标准的 Gamma 函数, $n_j^{(w)}$ 表示将词汇 w 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. 因为 $P(w|T)$ 可以近似为一系列 $P(w|z)$ 的调和平均值, 所以按下式求取其值:

$$\frac{1}{P(w|T)}=\frac{1}{M}\sum_{m=1}^{M}\frac{1}{P_m(w|z)} \tag{10}$$

实验结果如图 1 所示.

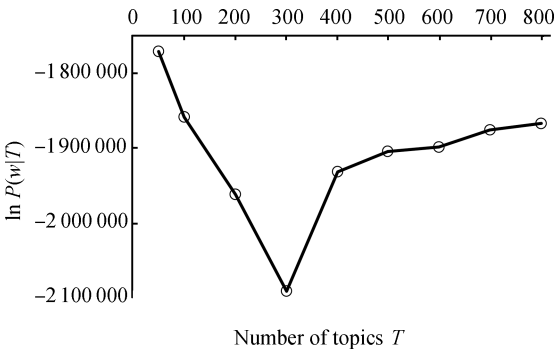

图 1　$\ln P(w|T)$ 与主题数目的关系

Fig. 1　The log-likelihood of the data for different settings of the number of topics T

由图 1 可以看出, 当主题数目 T 为 300 时, $\ln P(w|T)$ 的值最小, 随后开始急剧增大, 说明主题数目为 300 时, 模型对于语料库数据中有效信息的拟合最佳, 因此, 后续实验的主题数目取为 300.

4.3　Burn-in 及 Thinning 间距的选择

为了科学地确定 Burn-in (b) 值和 Thinning (c) 值, 本实验取 $T=300$, 以 4 个不同的初始值运行 Gibbs 算法, 若 b,c 的取值合适, 则抽样结果 ($\ln P(w|z)$ 的值) 随初始值的变化很小, 也可以说独立于初始值. 实验结果如图 2 所示.

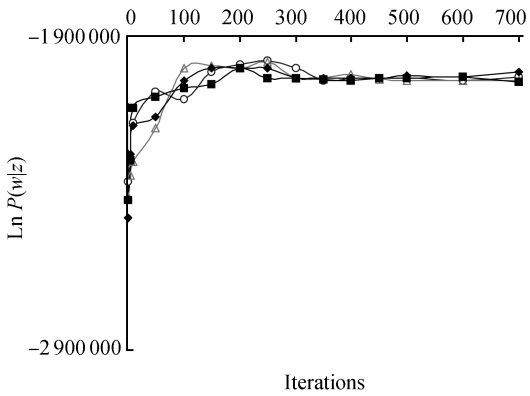

图 2　迭代数百次后 $\ln P(w|T)$ 趋于稳定

Fig. 2　The log-likelihood stabilizes after a few hundred iterations

从图 2 中可以看出, $\ln P(w|z)$ 的值在迭代数百次后稳定, 因此本文实验取 Burn-in 间距为 350, Thinning 间距为 50.

4.4 文本分割的测试

4.4.1 测试集

本实验利用 1997 年 3 月份人民日报手工标注的语料库构建 4 个测试集 T_{3-11}, T_{3-5}, T_{6-8}, T_{9-11}, T_{x-y} 表示所含主题片段的句数在 x 和 y 之间. 每一个测试集包括若干伪文本, 即由不同类的文本连接而成的形式上的文本, 要求相邻段落务必来自不同的类. 其所含的主题数平均为 7, 具体如表 1 所示.

表 1 实验中的测试集

Table 1 Test sets used in the experiments

	T_{3-11}	T_{3-5}	T_{6-8}	T_{9-11}
片段句数	3–11	3–5	6–8	9–11
伪文本数	109	127	115	98

4.4.2 度量标准

文本分割一般基于 P_k[12] 进行度量, 但从理论上讲 WindowDiff[13] 更为科学. 为了便于同类算法向前和向后的对比, 本文采用 P_k 和 WindowDiff 两种标准分别度量.

$$P_k = P(seg)P(miss) + (1 - P(seg))P(falsealarm) \quad (11)$$

$P(seg)$ 是指距离为 k 的两个句子分属不同主题片段的概率, 而 $1 - P(seg)$ 就是指距离为 k 的两个句子属于同一主题片段的概率, 本实验将两个先验概率取等值, 即 $P(seg) = 0.5$, $P(miss)$ 是算法分割结果缺少一个片段的概率, $P(falsealarm)$ 是算法分割结果添加一个片段的概率.

$$\text{WindowDiff}\,(ref, hyp) = \frac{1}{N-k}\sum_{i=1}^{N-k}\left(|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| > 0\right) \quad (12)$$

$b(i, j)$ 表示整句 s_i 和整句 s_j 间的边界数量, N 表示文本中的整句数量, k 取真实片段平均长度的一半, ref 代表真实分割, hyp 代表算法分割.

4.4.3 实验结果

Gibbs 抽样的主题数目 $T = 300$, 超参数 $\alpha = 50/T$, $\chi = 0.01$, 取 10 个不同的初始值运行算法, 每个初始值迭代 350 次, 然后每隔 50 次取一次样本, 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. 作为与本文方法的对比, 取 PLSA[5], LSA (Latent semantic analysis)[14], MDA (Multiple discriminant analysis)[15] 三种算法在 T_{3-11}, T_{3-5}, T_{6-8}, T_{9-11} 上进行测试, 结果如表 2. 可见, 本文算法对于 4 个测试集均有较好的表现, 而且实验表明测试结果比较稳定, 不同样本间的差别较小. 据文献 [16], 基于 PLSA 模型的分割, 其结果的随机性较大, 随迭代次数及主题数目的变化难于确定. 表中同时给出 P_k 和 WindowDiff 的值也验证了 P_k 的一些缺陷[13], 比如对于不同的片段长度, P_k 的评价准则不一致. 同样的错误在较小片段内给予较大的罚分, 而到了较大片段则给予较小的罚分或者忽略不计. 表中两个 11.94 的 P_k 值, 所对应的 WindowDiff 值差别极大, 也正说明了这一点. 由于 WindowDiff 对于错误的衡量比 P_k 更为精细, 所以基于同样的分割方法及测试集, 前者明显高于后者.

表 2 与 PLSA, LSA 以及动态规划的对比结果

Table 2 Segmentation results compared to PLSA, LSA, and dynamic programming

	T_{3-11} (%)	T_{3-5} (%)	T_{6-8} (%)	T_{9-11} (%)
本文算法	7.55 (16.12)	12.29 (27.03)	5.72 (18.18)	10.67 (24.32)
PLSA	16.79 (43.64)	13.81 (36.34)	13.26 (37.50)	11.94 (45.90)
LSA	13.12 (32.61)	15.21 (30.84)	10.02 (19.76)	12.17 (59.44)
MDA	11.61 (18.45)	11.38 (27.22)	11.94 (26.00)	11.00 (31.98)

4.5 主题提取的测试

4.5.1 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 10 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 (Institute of computing technology, Chinese lexical analysis system) 对其进行处理, 然后凭直觉给每个类以一定数目的标识词 (至多 5 个), 如下表 3 所示.

表 3 类及其标识词

Table 3 Categories and their identification words

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

4.5.2 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 (Precision) 定义如下:

$$precision = \frac{n_{\text{correct}}}{n_{\text{total}}} \quad (13)$$

其中, n_{correct} 指正确提取主题词的文本数, n_{total} 指测试文本的总数.

4.5.3 片段主题提取

以类为单位进行测试, 每个类取大约 100 个主题片段, 其测试集合如表 4 所示.

表 4 测试集及所包含的片段数目

Table 4 Test sets and their topics

类别	环境	经济	交通	教育	体育	政治	军事	计算机	艺术	医药
片段数	101	113	102	100	105	105	97	95	114	106

将本文方法与 TF-IDF[17] 及 Z-SCORE[17] 方法进行对比. TF-IDF 的计算方法为

$$weight_w(s) = \frac{tf_w(s) \times \ln\left(\dfrac{N}{n_w}\right)}{\sqrt{\sum\limits_{i=1}^{n}(tf_w(s))^2 \times \ln^2\left(\dfrac{N}{n_w}\right)}} \quad (14)$$

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

$$weight_w(s)=\frac{tf_w(s)-\frac{\sum_{i=1}^{N}f_i(w)}{N}}{\sqrt{\sum_{i=1}^{N}\left[f_i(w)-\frac{\sum_{i=1}^{N}f_i(w)}{N}\right]^2\times\frac{f_i(w)}{N}}} \tag{15}$$

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 和 7 时, 实验结果分别如表 5 和表 6. 从表中可见, 本文方法的结果在两种情况下均远远好于其他两种方法, 主要原因在于充分利用背景语料库的知识, 使主题词产生联想, 以此挖掘出隐藏于文本之中的内涵.

为了使得对比结果更为清晰, 图 3 绘出本文方法与 TF-IDF 以及 Z-SCORE 在主题词数为 5, 7, 10 时的对比情况, 每种方法取主题词数固定下的 10 个测试集结果的平均值.

从图 3 可以看出, 随着允许提取主题数目的增多, 三种方法的准确率均有提高, 但本文方法的变化幅度最小, 尤其当主题词数增加到 7 以后, 基本保持不变了, 这说明本文方法提取的主题词的先后顺序对于片段核心内容的反映程度比其他两种方法好很多.

表 5　主题词数为 5 的提取结果

Table 5　Subtopic identification results when the number of topic words is 5

类别	本文方法 (%)	TF-IDF (%)	Z-SCORE (%)
环境	81.29	47.17	9.44
经济	90.47	41.11	29.68
军事	72.09	50.52	10.17
计算机	75.46	56.37	24.33
交通	100.00	79.69	33.12
教育	100.00	80.00	21.98
体育	98.45	54.69	76.65
艺术	96.00	54.72	6.64
医药	93.27	61.67	44.83
政治	92.99	62.13	30.10

表 6　主题词数为 7 的提取结果

Table 6　Subtopic identification results when the number of topic words is 7

类别	本文方法 (%)	TF-IDF (%)	Z-SCORE (%)
环境	88.33	64.81	31.48
经济	95.54	50.00	34.62
军事	86.65	63.54	29.17
计算机	84.38	68.00	47.91
交通	100.00	96.92	81.54
教育	100.00	88.24	56.86
体育	98.99	72.31	89.23
艺术	99.86	70.37	16.67
医药	95.42	78.69	55.74
政治	97.23	76.36	49.09

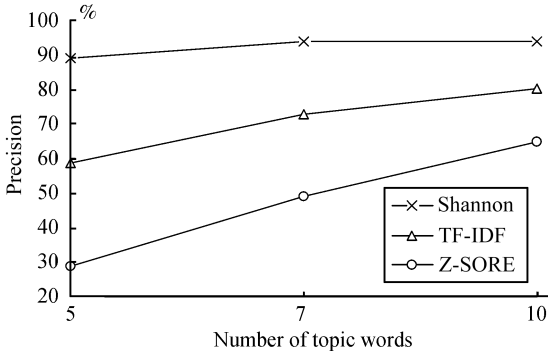

图 3　本文方法与 TF-IDF 以及 Z-SCORE 的结果对比

Fig. 3　Results of the method of this paper compared with those of TF-IDF and Z-SCORE

4.5.4　中心主题获取

令 $P(s)$ 为片段主题词元素的出现频率, 取 $P(s)>\mu$ 的片段主题词元素为中心主题词, 随着 μ 的提高, 准确率不断降低, 片段主题词数为 7 时, 其中心主题词获取结果如表 7 所示 (中心主题词数不限, 但少于 7).

表 7　μ 的取值及相应的中心主题提取的准确率

Table 7　μ and the precision of extraction of central topic

μ	0.5	0.6	0.7	0.8
准确率 (%)	98.78	97.25	88.41	74.54

5　相关研究对比

本文探讨适合主题分析的模型和方法, 将主题分割与主题识别集成在 LDA 框架下实现. 与本文研究最相关的工作是 STM (Statistical topic model)[4], MDA[14] 因涉及国内对于主题分割部分的研究, 所以也进行对比分析.

STM 是一种有限混合模型, 原则上讲, 这种模型假定一个文档仅仅呈现一个主题, 往往无法准确描述语料库及文档建模所需的数据信息, 同时, 由于没有对主题概率及词汇概率作任何假设, 导致局部极大值、过度拟合以及收敛速度过慢等问题. 本文作者在实验中发现, 基于该模型的主题分割错误率较高, 基本在 50% 左右, 主题词提取的准确率低于 TF-IDF 方法. 分析原因, 除了上述模型自身存在的问题外, 还由于对模型参数的估算基于单一文档的部分信息 (包括 h 句的块), 而非语料库丰富的知识, 但毕竟一个块内提供的信息过于有限, 所以无法准确估算参数值.

MDA 方法定义了 4 种全局评价函数, 寻找满足分割单元内距离最小化和分割单元间距离最大化条件的最好分割方式, 实现对文本分割模式的全局评价. 其优点在于通用性强, 无需语料库, 缺点是片段边界的确定仅仅依赖本文档的内部信息, 难于实现更好的分割. 表 2 的实验结果同样说明, 采用 MDA 方法, 其分割错误率 (P_k) 极为集中 (11% 左右). 而本文方法更多地依赖于语料库的训练, 因此当语料库信息充分, 测试文档与训练语料结构类似时就会呈现更好的分割效果 ($P_k=5.72\%$).

6　结语

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结构 – 文本分割之上提取片段主题词并总结全文的中心主题词. 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准确性, 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. 实验结果表明, 本文方法有很好的分析表现, 可以为文本推理的研究提供坚实的基础. 下一步的工作将尝试采用文献 [18] 的方法提高模型训练的速度, 使得该方法可行.

References

1 Kehagias A, Nicolaou A, Petridis V, Fragkou P. Text segmentation by product partition models and dynamic programming. *Mathematical and Computer Modeling*, 2004, **39**(2-3): 209−217

2 Gina-Anne L. Prosody-based topic segmentation for mandarin broadcast news. In: Proceedings of the 9th American Chapter of the Association for Computational Linguistics-Human Language Technologies. Boston, USA: Association for Computational Linguistics, 2004. 137−140

3 Olivier F. Using collocations for topic segmentation and link detection. In: Proceedings of the 19th International Conference on Computational Linguistics. Taipei, China: Association for Computational Linguistics, 2002. 1−7

4 Li H, Yamanishi K. Topic analysis using a finite mixture model. *Information Processing and Management*, 2003, **39**(4): 521−541

5 Hofmann T. Probabilistic latent semantic analysis. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Stockholm, Sweden: Morgan Kaufmann, 1999. 289−296

6 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003, **3**: 993−1022

7 Steyvers M, Griffiths T. Probabilistic topic models. *Handbook of Latent Semantic Analysis*. New Jersey: Springer, 2007

8 Minka T, Lafferty J. Expectation-propagation for the generative aspect model. In: Proceedings of the 18th Uncertainty in Artificial Intelligence. Alberta, Canada: Morgan Kaufmann, 2002. 352−359

9 Griffiths T L, Steyvers M. Finding scientific topics. In: Proceedings of the National Academy of Sciences. USA: Springer, 2004. 5228−5235

10 Heinrich G. Parameter Estimation for Text Analysis, Technical Report, University of Leipzig, Germany, 2008

11 Brants T, Chen F, Tsochantaridis I. Topic-based document segmentation with probabilistic latent semantic analysis. In: Proceedings of the 11th International Conference on Information and Knowledge Management. McLean, USA: ACM, 2002. 211−218

12 Beeferman D, Berger A, Lafferty J. Statistical models for text segmentation. *Machine Learning*, 1999, **34**(1-3): 177−210

13 Pevzner L, Hearst M A. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 2002, **28**(1): 19−36

14 Choi F Y Y, Wiemer-Hastings P, Moore J D. Latent semantic analysis for text segmentation. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. Pittsburgh, USA: Camegie Mellon University, 2001. 109−117

15 Zhu Jing-Bo, Ye Na, Luo Hai-Tao. Text segmentation model based on multiple discriminant analysis. *Journal of Software*, 2007, **18**(3): 555−564
(朱靖波, 叶娜, 罗海涛. 基于多元判别分析的文本分割模型. 软件学报, 2007, **18**(3): 555−564)

16 Shi Jing, Dai Guo-Zhong. Text segmentation based on PLSA model. *Journal of Computer Research and Development*, 2007, **44**(2): 242−248
(石晶, 戴国忠. 基于 PLSA 模型的文本分割. 计算机研究与发展, 2007, **44**(2): 242−248)

17 Liu Y, Ciliax B J, Borges K, Dasigi V, Ram A, Navathe S B. Comparison of two schemes for automatic keyword extraction from MEDLINE for functional gene clustering. In: Proceedings of the IEEE Computational Systems Bioinformatics Conference. Washington D. C., USA: IEEE, 2004. 394−404

18 Porteous I, Newman D, Ihler A, Asuncion A, Smyth P, Welling M. Fast collapsed Gibbs sampling for latent dirichlet allocation. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA: ACM, 2008. 569−577

石 晶 长春工业大学讲师. 2007 年获得中国科学院软件研究所博士学位. 主要研究方向为语言与信息处理. 本文通信作者.
E-mail: crystal1087@126.com
(**SHI Jing** Lecturer at Changchun University of Technology. She received her Ph. D. degree from the Institute of Software, Chinese Academy of Sciences in 2007. Her research interest covers processing of natural language and information. Corresponding author of this paper.)

范 猛 长春工业大学副教授. 主要研究方向为材料科学.
E-mail: fanmeng@mail.ccut.edu.cn
(**FAN Meng** Associate professor in the Department of Science and Research Administration, Changchun University of Technology. His main research interest is science of material.)

李万龙 长春工业大学教授. 主要研究方向为知识工程.
E-mail: lwl@mail.ccut.edu.cn
(**LI Wan-Long** Professor at Changchun University of Technology. His main research interest is knowledge engineering.)