采用并行遗传算法的文本分割研究

赵煜[1]，蔡皖东[1]，樊娜[1]，刘念[2]
（1. 西北工业大学计算机学院，710072，西安；2. 西安建筑科技大学图书馆，710055，西安）

摘要：针对短篇幅文本数据稀疏的特性，提出了一种利用外部语料库知识提高短篇幅文本分割准确率的方法. 该方法分2个步骤完成：① 利用Gibbs采样方法估计语料库对应的潜在狄利克雷分配(LDA)模型，并利用该模型推断目标文本的潜在语义结构信息；② 通过定义语义段落内凝聚性和语义段落间发散性2个目标函数，将文本分割问题转化为多目标优化问题. 采用一种针对文本分割的并行遗传算法，获得全局最优解. 通过实验，在文本数据稀疏的情况下，该算法在准确率方面优于多元判别分析(MDA)方法和基于LDA的文本分割方法，对于提高文本分割的准确率是可行和有效的.

关键词：中文信息处理；文本分割；遗传算法
中图分类号：TP39 **文献标志码**：A **文章编号**：0253-987X(2009)12-0040-05

Research of Text Segmentation Based on Parallel Genetic Algorithm

ZHAO Yu[1], CAI Wandong[1], FAN Na[1], LIU Nian[2]
(1. College of Computer Science, Northwestern Polytechnical University, Xi′an 710072, China;
2. Library, Xi′an University of Architecture and Technology, Xi′an 710055, China)

Abstract: Focusing on the data sparseness of short texts, an algorithm based on knowledge from external corpus is proposed to improve the accuracy of text segmentation, which contains two steps: Gibbs sampling is adopted to estimate the LDA model; corresponding to the corpus and the latent semantic structure information of the text is inferred based on the LDA model. Two objective functions of internal cohesion and external dissimilarity are then defined to transform text segmentation into a multi-objective optimization problem. A parallel genetic algorithm based on the objective functions is emloyed to obtain the global optimal solution for text segmentation. According to the experiments, the proposed algorithm achieves higher accuracy than the MDA and LDA-based methods in the case of data sparseness.

Keywords: Chinese information processing; text segmentation; genetic algorithm

文本分割技术按照文本内容的语义相关度来识别独立意义单元之间的边界，获取文本子主题结构. 这些信息的使用将有助于提高问答式信息检索、文本文摘、细粒度文本情感分析等信息获取技术的性能. 因此，文本分割作为自然语言理解领域的一种预处理技术具有重要的研究意义.

目前的主流文本分割方法主要利用词频计算句子之间的相似度[1]，进而实现对文本的分割. 在这类方法中，文本分割策略分为两种：其一，局部最优策略，即仅依据相邻片段之间的相似度来确定文本分割的边界，分割阈值预先确定或通过局部信息计算获得，分割结果的随机性大、准确率较低；其二，全局最优策略，即考虑文本中所有片断之间的相似度，采用动态规划、遗传算法实现文本分割，其分割准确率更高. 随着概率潜在语义分析(PLSA)模型、LDA模型在自然语言处理领域中的应用，人们提出了利用

收稿日期：2009-06-22. 作者简介：赵煜(1981—)，男，博士生；蔡皖东(联系人)，男，教授，博士生导师. 基金项目：国家“863计划”资助项目(2009AA01Z424)；教育部高等学校博士学科点专项科研基金资助项目(200806990030).

概率统计模型分析结果推断分割边界的方法. 这类方法通过挖掘文本中隐含的语义结构，利用这些信息进行文本分割[2-4]. 由于这类方法依赖于训练语料库，因此当语料库信息充分、测试文档与训练语料主题类似时，方法会呈现较好的分割效果.

以上分割方法都是利用词频[1]或词汇在分割基本单元①中的概率[2-4]来计算分割基本单元之间的相似度. 然而，由于一些网络文本（如论坛、博客、网络聊天消息）篇幅较短，数据稀疏，无法为分割基本单元相似度计算提供足够的词汇共现信息，因此主流分割方法在短文本分割中无法达到理想效果. 针对这个问题，本文提出一种适用于文本分割的并行遗传算法. 这种方法除了利用文本本身信息，还考虑了与文本主题相关的语料库信息，依据相关语料库的 LDA 模型，提取待分割文本隐含的语义结构信息，进行分割基本单元相似度计算，同时发挥遗传算法的全局优化能力，寻找全局最优解，提高文本分割的性能.

1　文本分割系统模型

由于 PLSA 模型中的文本概率值与训练文本集合密切相关，导致 PLSA 模型处理新文本的能力较差，同时模型参数数量随文本数量增加而线性增长. LDA 模型[5]克服了 PLSA 模型存在的问题，是一个完全的生成模型. 因此，本文采用 LDA 模型，并采用 Gibbs 采样[6]估计 LDA 模型的参数，推断待分割文本的相关语义结构信息.

本文提出的文本分割方法分为 4 个步骤，系统模型如图 1 所示.

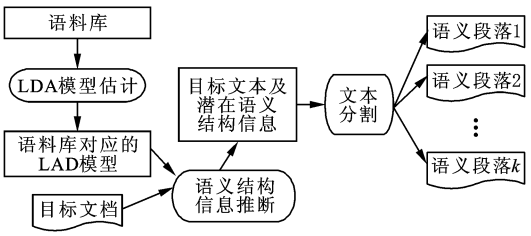

图 1　文本分割系统模型

（1）构建语料库. 利用网络蜘蛛完成语料搜集工作，并通过文本聚类，按照主题构建语料库. 语料库为文本语义结构挖掘提供了更多依据，文本与语料库之间的主题相关度越高，文本的语义结构描述就越准确.

（2）采用 Gibbs 采样方法，估计语料库对应的 LDA 模型.

（3）推断潜在语义结构信息. 根据主题选择不同的语料库，依据语料库训练结果，挖掘待分割文本的语义结构信息. 文本的语义结构信息包括特定句子中词汇所属子主题类型及词汇在句子中的概率.

（4）文本分割. 利用语义结构特征，计算句子之间的相似度，并采用遗传算法进行文本分割.

2　基于并行遗传算法的文本分割法

在文本分割任务中，需要同时考虑语义段落内凝聚性和语义段落间的发散性. 由于两个属性相互制约，难以同时达到最优值，所以文本分割可看作一个多目标优化问题. 借鉴 Zitzler 解决多目标优化问题的方法[7]，本文提出的针对文本分割问题的并行遗传算法如图 2 所示.

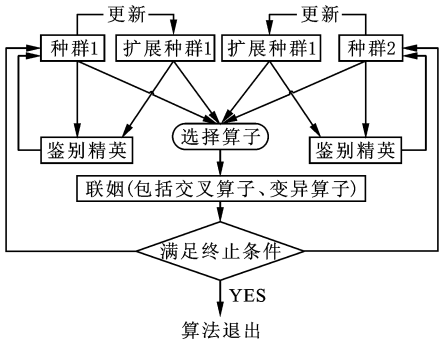

图 2　针对文本分割问题的并行遗传算法

2.1　编码方案

文本分割结果由句子对应的分割标记描述. 当分割标记为 0 时，句子与上文语义相近；当分割标记为 1 时，该句子开启新的语义段落是文本分割的边界. 文本分割结果是一个二进制符号串，因此二进制编码方法适合文本分割问题和结果的表述.

2.2　种群初始化

通过文本分割实验发现：为了保证论述的连贯性，文本中存在一些与前后文相关性不强的句子，称之为过渡句. 如果将这些句子分割为独立的语义单元，将不能表达一个完整的主题意义. 基于这个考虑，在种群初始化过程中，本文根据实验经验对随机数生成方法产生的个体进行过滤. 过滤标准包括语义段落的最小长度，文本包含语义段落的最小数量.

2.3 适应度函数

定义文本 $D=s_1 s_2 \cdots s_k \cdots s_N$ 表示文本中包含的

① 分割基本单元可以是句子、段落等，本文研究基于句子的文本线性分割问题.

句子总数，$\boldsymbol{s}_k$ 为句子 k 对应的潜在语义结构信息向量，$\boldsymbol{s}_k=s_{k1}\,s_{k2}\cdots s_{kj}\cdots s_{kT}$ 表示出句子 k 中词汇属于子主题 j 的频率. 因此，可以将文本分割过程转化为序列型数据空间划分过程. 划分区域所包含成员之间的平均距离越小，对应的质心之间的平均距离越大，划分结果越趋向于最优解，前者对应语义段落内部的凝聚性，后者对应语义段落之间的发散性. 所以，利用这两个属性定义种群中个体的适应度函数.

(1)语义段落内凝聚性. 文本定义同上文，$B=b_1b_2\cdots b_j$ 表示文本 D 的一个分割结果，j 表示语义段落的数量，则语义段落内的凝聚性定义为

$$C_{\mathrm{coh}}=1-\sum_{n=1}^{j}\frac{1}{k}\sum_{s_i\in b_n}\sum_{l=1}^{T}(s_{il}-a_{nl})^2 \tag{1}$$

$$a_{nl}=\frac{1}{|\,b_n\,|}\sum_{s_i\in b_n}s_{il};\ \boldsymbol{a}_n=[a_{n1},a_{n2},\cdots,a_{nl},\cdots,a_{nT}] \tag{2}$$

式中：$|b_n|$表示第 n 个语义段落中包含的句子数；$\boldsymbol{a}_n$ 表示语义段落对应的平均向量，其中 a_{nl} 是该向量的第 l 个分量.

(2)语义段落间发散性. 根据语义段落内凝聚性的符号表示，定义文本 D 的语义结构信息平均向量 $\boldsymbol{c}$ 为

$$c_l=\frac{1}{k}\sum_{i=1}^{k}s_{il};\quad \boldsymbol{c}=[c_1,c_2,\cdots,c_T] \tag{3}$$

根据式(3)，语义段落间发散性定义为

$$D_{\mathrm{is}}=\sum_{n=1}^{j}\frac{|\,b_n\,|}{k}\sum_{l=1}^{T}(a_{nl}-c_l)^2 \tag{4}$$

(3)适应度函数的定义. 借鉴文献[7]解决多目标优化问题的方法，结合上文定义的两个目标函数，本文定义了种群中个体的适应度函数. 如果 P_t 表示遗传算法中的种群，$\overline{P}_t$表示扩展种群，则个体的适应度函数为

$$F(x_i)=\begin{cases}\{|\,x_j\mid x_j\in P_t\wedge C_{\mathrm{coh}}(x_i)\geqslant C_{\mathrm{coh}}(x_j)\wedge D_{\mathrm{is}}(x_i)\geqslant D_{\mathrm{is}}(x_j)\}\,|\,/(|\,P_t\,|+1) & x_i\in\overline{P}_t\\ 1+\sum\limits_{x_j\in\overline{P}_t\wedge C_{\mathrm{coh}}(x_j)\geqslant C_{\mathrm{coh}}(x_i)\wedge D_{\mathrm{is}}(x_j)\geqslant D_{\mathrm{is}}(x_i)}F(x_j) & x_i\in P_t\end{cases} \tag{5}$$

2.4 遗传算子

(1)选择算子. 在种群选择过程中，本文首先采用精英保留策略，保留 P_t 和$\overline{P}_t$ 中的精英个体，直接进入下一代种群. 然后采用轮盘赌方法，分别从 P_t 和 $\overline{P}_t$ 中选择个体，比较两个体的适应度，选择适应度小的个体进行交叉和变异操作.

(2)交叉算子. 本文采用单点交叉，即依据概率决定个体对是否进行联姻，并随机产生交叉位置，交换两个配对个体的部分基因组. 为了防止近亲繁殖，参与交叉的个体必须属于不同的种群，并且只有当个体间汉明距离超过一定阈值时，才允许在二者之间进行交叉操作.

(3)变异算子. 在遗传算法中，变异算子能够维持种群的多样性，较优个体部分基因组的调整还可能使个体更加逼近最优解. 但是，过大的变异概率也会导致过多的随机搜索，降低算法效率. 因此，本文采用一种自适应的变异算子，用种群相似度衡量种群的一致性，当种群中个体趋于一致时，提高个体的变异概率.

定义种群 P 中任意 2 个体 x_i、x_j 的相似度为

$$S_{\mathrm{im}}(x_i,x_j)=\sum_{l=1}^{T}\sim(x_{i,l}\oplus x_{j,l})/T \tag{6}$$

根据式(6)，种群 P 的相似度定义为

$$S_{\mathrm{im}}(P)=\frac{2\times\sum\limits_{i\neq j\wedge x_i,x_j\in P}S_{\mathrm{im}}(x_i,x_j)}{|\,P\,|\,(|\,P\,|-1)} \tag{7}$$

当 $S_{\mathrm{im}}(P)\to1$ 时，说明种群的相似程度很高，此时需提高个体的变异概率；当 $S_{\mathrm{im}}(P)\ll1$ 时，表示种群中个体在解空间中分布均匀，应设定小的变异概率.

本文采用两种变异策略：①改变分割边界位置；②当分割边界位置改变造成个体不满足文本分割要求时，生成新的个体进行替换. 两种变异策略相互配合，保证算法具有良好的搜索性能.

2.5 种群及扩展种群的更新

为了保证优秀个体参与下一轮进化过程，在种群更新过程中，只有新生个体的适应度超过父代个体，新生个体才能替换父代个体进入下一代种群. 同时，为保持种群个体的多样性，算法采用小生境技术[8]，新生个体仅替换与之相似的父代个体，其中根据式(6)计算个体之间的相似度.

扩展种群用于保存与最优解较接近的个体，当算法满足终止条件时，文本分割结果也从扩展种群中获得. 扩展种群的更新过程如图 3 所示.

3 实 验

3.1 实验过程

目前，文本分割结果的评价通常采用两种方法：一是人为将不同内容的文本连接起来，文本内容的变化指示语义片段的边界；二是使用正常文本，按照人的判断对算法结果进行评价. 由于评价人数、判断

输入:P_t遗传算法中第t代种群
　　$\overline{P}_{t-1}$遗传算法中第$t-1$代种群
输出:$\overline{P}_t$遗传算法中t代扩展种群
过程:
(1) $P'=\overline{P}_{t-1}$
(2) $P'=P'+\{x_i|x_i\in P_t\wedge\nexists x_j, C_{\text{coh}}(x_j)\geqslant C_{\text{coh}}(x_i)$
　　$D_{\text{is}}(x_j)\geqslant D_{\text{is}}(x_i), x_j\in P_t\}$
(3) if $\exists x_i, x_j\in P'$
　　$C_{\text{coh}}(x_j)\geqslant C_{\text{coh}}(x_i)\wedge D_{\text{is}}(x_j)\geqslant D_{\text{is}}(x_i)$
　　then $P'=P'-\{x_i\}$
(4) $\overline{P}_t=P'$

图 3　扩展种群的更新过程

水平的不同可能会造成人工评价结果存在差异,进而影响评价的客观性,因此本采用第一种评价策略.

目前还没有一个公开、通用的文本分割测评语料.《人民日报》作为中文信息处理实验中常用的语料源,数据稳定性好,文献[1-2]都选取《人民日报》作为实验语料,因此本文也采用该语料源,保证实验数据相似. 本文语料库规模为 2 000 篇,体裁和内容主要来自时事评论领域,同时利用哈尔滨工业大学的语言技术平台(LTP)对数据集进行文本预处理,并依据知网构造停用词表. 本文实验构造了 2 个测试集,每个测试集中包含若干伪文本,伪文本包含 20～30 个语义段落,语义段落的长度在特定范围内,具体情况如表 1 所示.

表 1　实验中的测试集

测试集	语义段落的长度	伪文本数量
T1	2～4	50
T2	2～9	50

由于 LDA 模型参数估计过程对子主题数目非常敏感,因此确定子主题数目是 LDA 模型估计的关键. 文献[2,6]依据 $\lg P(w|T)$峰值确定子话题数目,确定过程没有考虑主题信息. 与文献[2]不同,本文首先按照主题构建语料库,并对每个主题语料库进行层次聚类[9-10]. 结果中簇的个数即为子主题数目. 由于子主题数目是通过聚类特定主题语料库获得,有助于减少不相关主题信息对 LDA 模型估计的影响,进而提高文本分割的精度. Gibbs 采样相关参数借鉴文献[2]方法确定.

3.2　评价方法

本文首先采用传统的正确率、召回率以及 F 值来评价文本分割算法的性能. 同时,由于传统的评价标准只考虑绝对匹配的结果,不能全面地评价文本分割的效果,因此本文还采用 WindowDiff[11] 和 P_μ[12] 两种评价标准,上述评价标准的定义为

$$W_{\text{WindowDiff}}(r_{\text{ef}},h_{\text{yp}})=\frac{1}{N-m}\cdot\sum_{i=1}^{N-m}(|b(r_{\text{ef}_i},r_{\text{ef}_{i+m}})-b(h_{\text{yp}_i},h_{\text{yp}_{i+m}})|>0)\quad(8)$$

式中:$b(i,j)$表示句子 i 与句子 j 之间包含的边界数量;N 表示文本包含句子总数;m 表示语义段落平均长度的一半;r_{ef}表示标准解;h_{yp}表示算法得出的结果.

另一评价标准

$$P_\mu(r_{\text{ef}_i},h_{\text{yp}_j})=\sum_{1\leqslant i\leqslant j\leqslant N}D_\mu(i,j)\delta(\gamma_{\text{ef}_i},\gamma_{\text{ef}_j})\overline{\oplus}\,\delta(h_{\text{yp}_i},h_{\text{yp}_j})\quad(9)$$

式中:$\delta(i,j)$是指示函数,句子 i、j 属于同一个语义段落,函数取值为 1,反之为 0;D_μ 为随机选取的句子对在文本中的概率

$$D_\mu(i,j)=r_\mu e^{-\mu|i-j|}\quad(10)$$

其中:r_μ 为归一因子;μ 为语义段落长度的均值.

3.3　算法策略及性能分析实验

在本文算法执行中,有 2 个阶段可能采用不同的策略,这 2 个阶段是:①选择文本分割使用的语义结构特征;②选择种群个体变异的方案. 具体策略如表 2 所示. 为了分析各种策略方案对算法性能的影响,本文分别采用以上策略进行实验,实验结果如表 3 所示.

表 2　本文算法涉及的策略

算法执行阶段	策略	记号
语义结构特征选择	词汇在句子中的概率	Ws
	词汇所属子主题类型	Wt
变异方案选择	考虑分割限制	Il
	不考虑分割限制	Nil

表 3　算法策略分析实验结果

阶段	测试集	策略	P_r	R_c	F	$W_{\text{WindoDiff}}$	P_μ
话题信息选择策略	T1	Wt	0.78	0.88	0.82	0.30	0.87
		Ws	0.67	0.75	0.71	0.56	0.78
	T2	Wt	0.57	0.80	0.67	0.48	0.80
		Ws	0.47	0.81	0.61	0.62	0.72
变异方案选择策略	T1	Il	0.78	0.88	0.82	0.30	0.87
		Nil	0.36	0.88	0.53	0.78	0.46
	T2	Il	0.57	0.80	0.67	0.48	0.80
		Nil	0.24	0.89	0.36	0.92	0.30

从表3可以看出,采用子主题类型信息的算法比采用词汇在句子中出现概率的算法取得了更好的文本分割效果,个体变异操作中考虑文本分割限制,促使算法向最优解方向收敛.

3.4 对比实验

文献[1]同时考虑语义段落内距离、语义段落之间距离和语义段落长度等特征进行实验,提高了文本分割的性能.文献[2]将余弦度量方法与动态常数边界识别方法进行组合,在4组测试样本中,2组结果达到最优.因此,本文选取文献[1-2]中的这2种方法与本文方案进行比较,对比实验结果如表4所示.

表4 对比实验结果

测试集	方案	P_r	R_c	F	$W_{WindoDiff}$	P_μ
T1	本文方法	0.78	0.88	0.82	0.30	0.87
	MDA	0.42	0.63	0.50	0.52	0.58
	LDA-b	0.38	0.38	0.38	0.30	0.58
T2	本文方法	0.57	0.80	0.67	0.48	0.80
	MDA	0.41	0.75	0.52	0.64	0.65
	LDA-b	0.33	0.40	0.36	0.41	0.63

从表4可以看出,本文算法在准确率、召回率、F值、P_μ值等方面优于baseline方法,说明在数据稀疏情况下,该方法有效提高了文本分割的性能,同时证明全局优化有助于提高文本分割性能.

4 结 论

针对短篇幅文本存在数据稀疏的特性,本文深入研究了文本分割问题,定义了语义段落内凝聚性、语义段落间离散性两个目标函数,将文本分割问题转化为多目标优化问题,采用并行遗传算法进行求解.实验结果表明,考虑文本中潜在语义结构信息及采用全局优化方法有助于提高文本分割性能.

参考文献:

[1] 朱靖波,叶娜,罗海涛. 基于多元判别分析的文本分割模型[J]. 软件学报,2007,18(3):555-564.
ZHU Jingbo, YE Na, LUO Haitao. Text segmentation model based on multiple discriminant analysis [J]. Journal of Software, 2007, 18(3): 555-564.

[2] 石晶,胡明,石鑫,等. 基于LDA模型的文本分割[J]. 计算机学报,2008,31(10):1865-1873.
SHI Jing, HU Ming, SHI Xin, et al. Text segmentation based on model LDA [J]. Chinese Journal of Computers, 2008, 31(10): 1865-1873.

[3] BRANTS T, CHEN F, TSOCHANTARIDIS I. Topic-based document segmentation with probabilistic latent semantic analysis [C]// Proceedings of the 11th International Conference on Information and Knowledge Management. McLean, Virginia, USA: ACM Press, 2002: 211-218.

[4] 石晶,戴国忠. 基于PLSA模型的文本分割[J]. 计算机研究与发展,2007,44(2):242-248.
SHI Jing, DAI Guozhong. Text segmentation based on PLSA model [J]. Journal of Computer Research and Development, 2007, 44(2): 242-248.

[5] BLEI D M, ANDREW Y N, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[6] GRIFFITHS T, STEYVERS M. Finding scientific topics [C]// Proceedings of the National Academy of Sciences. Washington, USA: National Academy of Sciences, 2004: 5228-5235.

[7] ZITZLER E. Evolutionary algorithms for multiobjective optimization: methods and applications [D]. Zurich, Swiss: Swiss Federal Institute of Technology Zurich, 1999.

[8] 雷英杰,张善文,李续舞,等. 遗传算法工具箱及应用[M]. 西安:西安电子科技大学出版社,2005.

[9] WITTEN I H, FRANK E. Data mining: practical machine learning tools and techniques [M]. Singapore: Elsevier, 2005.

[10] JUNG Y. Design and evaluation of clustering criterion for optimal hierarchical agglomerative clustering [D]. Twin Cities, USA: University of Minnesota, 2001.

[11] PEVZNER L, HEARST M. A critique and improvement of an evaluation metric for text segmentation [J]. Computational Linguistics, 2002(28): 1-19.

[12] BEEFERMAN D, BERGER A, LAFFERTY J. Text segmentation using exponential models [C]// Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing. Rhode Island, USA: EMNLP, 1997: 35-46.

(编辑 杜秀杰)