LDA主题驱动的中文多文档自动文摘方法*

张明慧，王红玲，周国栋

（苏州大学 计算机科学与技术学院，江苏 苏州 215002；
江苏省计算机信息处理技术重点实验室，江苏 苏州 215002）

E-mail:{ 20084227065097,hlwang,gdzhou }@suda.edu.cn

摘 要：多文档自动文摘能够帮助人们自动、快速地获取信息，本文实现了一个基于主题模型的中文多文档自动文摘系统，其中主题模型采用浅层狄利赫雷分配(Latent Dirichlet Allocation, LDA)，该模型是一个多层的产生式概率模型，能够检测文档中的主题分布。该方法使用 LDA 为多文档集合建模，通过计算句子在不同主题上的概率分布之间的相似度作为句子的重要度，并根据句子重要度进行文摘句的抽取。实验结果表明，该方法所得到的文摘，性能优于传统的文摘方法。

关键词： 中文自动文摘；主题模型；LDA；多文档；

Chinese multi-document summarization based on LDA Topic-Oriented method

Zhang Ming-hui, Wang Hong-ling, Zhou Guo-dong

(School of Computer Science & Technology Soochow University，Suzhou 215002
Jiangsu Provincial Key Laboratory of Computer Information Processing Technology, Suzhou 215006, China)

E-mail:{ 20084227065097,hlwang,gdzhou }@suda.edu.cn

Abstract: Multi-document summarization can help people access to information automatically and fast. In this paper, we propose a new method for Chinese multi-document summarization based on LDA topic model. The LDA model is a multi-level generative probabilistic model, it can detect the topic distribution of the ducument. In the method, we model the document using LDA, and then calculate the distance between a sentence and the given multi-documents via their topic probability distributions as the weight of the sentence, finally, we extract sentences according to the weight of the sentence. Experiment results show that the performance is a clear superiority over the traditional method under the proposed evaluation scheme.

Key words: Automatic Document Summarization; Latent Dirichlet Allocation; Topic Model; Multi-document; Chinese

1 引言

文章摘要是指从大量的文本中提取出其中的重要信息来代表文本的中心思想，文档摘要可以帮助人们快速、高效地获取信息。随着网络的普及，网络上的信息量日益剧增，单篇文档摘要已经不能满足人们的需要，多文档自动文摘技术应运而生。和单文档自动文摘相比，多文档自动文摘需要考虑文档之间的相关性，以及文档信息之间的冗余性，因此在多文档自动文摘中，如何选择文摘句决定了文摘质量的好坏。本文将主体模型融入到中文的多文档自动文摘系统中，使用词包（bag of words）来表示主题，并应用LDA主题模型对多文档系统进行建模，提取文章的主题

*基金资助：国家自然科学基金(60673041, 60873150)；江苏省高校自然科学重大基础研究项目（08KJA520002）。
作者简介：张明慧（1986-），女，硕士研究生，主要研究方向：自然语言处理；王红玲（1975-），女，讲师，通讯作者，主要研究方向：自然语言处理；周国栋（1967-），男，教授，博士生导师，研究方向：自然语言处理

特征信息，设计了一种浅层语义分析的多文档自动摘要方法。

文章第2部分简述了自动文摘的相关研究。第3部分介绍了基于LDA主题模型的自动文摘系统，重点描述了系统LDA特征。第4部分对实验结果进行了分析和比较。最后第5部分对本文进行了总结，并对后期工作进行了展望。

2 相关研究

目前常用的多文档自动文摘方法有两种：一种是将单文档自动摘要的方法用到多文档自动摘要上，如：Radev[1]采用的基于信息抽取技术的多文档自动文摘系统，这种方法虽然取得了较好的性能，但是忽略了文档之间的相关性。另一种是利用多文档集合信息的方法，如：Boros[2]采用的基于子主题聚类的方法，这种方法将文档集合作为一个整体来研究，对文档集合中的句子聚类，分为不同的主题类，然后分别从不同的主题类别中抽取句子组成摘要，这是目前比较流行的主题驱动的多文档摘要方法，但是这种方法也存在一定的缺陷：由于受限于摘要长度，并不是所有的从主题类中选出的句子都能作为摘要内容，这样可能使得产生的文摘内容代表性不强。

使用主题模型进行英文自动文摘并不少见，国外相关的研究有Arora等（2008）[5]，Bhandari等 (2008)[7]。其中Arora等使用LDA为文档建模，将每个句子对应一个主题，将主题表示为单词的权重矩阵，使用SVD求解句子集合的正交表示，作为选择句子的依据，从而降低文摘信息的冗余度。得到的主题概率作为句子选择的特征，并提出了基于推论的、半生成型的和全生成型的三种句子选择形式，并且通过实验说明基于推论的方法效果最好。在国内，吴晓锋[8]等将LDA提取的主题作为特征加入CRF模型中进行训练，有效地提高以传统特征为输入的CRF文摘系统的质量。中文的多文档文摘研究工作还处于起步阶段，从已发表的研究来看，中文的自动文摘系统大多数是基于规则和统计的方法，采用词频，位置信息，与标题相似度等特征作为衡量句子权重的方法。基于主题驱动的中文自动文摘系统相对较少，国内使用主题驱动的自动文摘主要采用的是利用k-means主题分类及k-means改进的主题分类[3]的方法。

3 自动摘要系统

本文构建的基于 LDA 模型的中文多文档自动文摘系统分为如下四个部分：文档预处理，多文档 LDA 建模，句子权重计算，文摘句的抽取（参见图 1）。

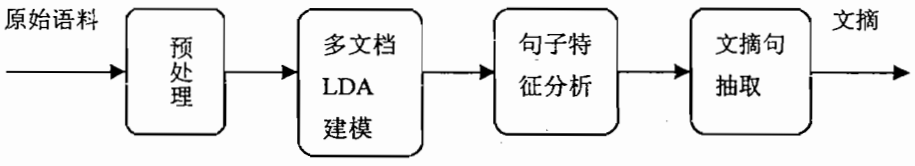

图 1　系统流程图

3.1 文档预处理

本文使用中科院的 ICTCLAS 分词系统对文档进行分词处理，然后在预处理阶段根据停用词表去除停用词，另外根据文档特征去掉了对文摘作用不大的介词、虚词、数词等词语，提高了系统准确率。

3.2 多文档 LDA 建模

LDA 是一个多层的产生式概率模型，包含词，主题，和文档三层结构。通过浅层的主题将词和文档关联起来。LDA 做了词袋假设（Bag of Words），即在模型中不考虑词汇的语法或顺序，只考虑它们出现的次数。模型将文本看成是有多个浅层的主题混合组成的，每个主题对应所有词汇上的一个多项式概率分布，这些主题被集合中的所有文档所共享，每个文档有一个特定的主题比例，即特定的主题分布。给定一个文档集合 D，包含 M 篇文档和 V 个不同的词汇。每个文档 d 包含一个词序列{w_1，w_2，……w_n}。在集合 D 对应的 LDA 模型中，假设主题数目固定为 K，则生成一篇文档 d 通过以下两个过程：

(1) Choose $\theta \sim Dir(\alpha)$

(2) For each of the W *words* w_i

a) Choose a topic $z_i \sim Multinomial(\theta)$

b) Choose a word w_n *from*, $p(w_i \mid z_i, \beta)$

其中 θ 是一个 1×k 的随机行向量，p(θ)是 θ 的分布，z_n是离散随机变量，在主题 T 中取 k 个离散值，p(z |θ)是给定 θ 时 z 的条件分布，w_n是离散随机变量，在词汇表 V 中取|V|个离散值，p(w | z)是给定 z_n时 w 的条件分布，可以把它看作 k×|V|的矩阵。LDA 产生一篇文档 d 前，先根据 Dirichlet 分布随机生成一个 1*k 的向量 θ，然后根据分布 p(z | θ)随机选取 p(w | z)的第 z_i 行，接着是根据分布 p(w |z = z_i)随机选取 z_i 行的第 w_i 列，从而来产生文档 d 中的所有单词。

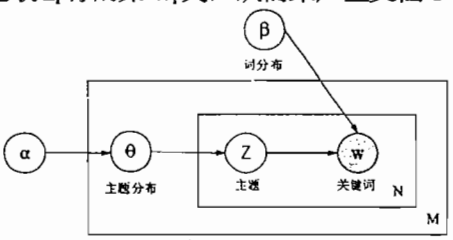

图 2 LDA 概率模型图

图 2 是 LDA 生成过程的概率模型图。空心点表示隐含变量，实心点表示可观察值；矩形表示重复过程，即词袋（Bag of Words）。外层矩形表示从 Dirichlet 分布中为文档集 D 中的每个文档 d 反复抽取主题分布 θ_d；内层矩形表示从主题分布中反复抽样产生文档 d 的词{w_1，w_2，…… w_n}。我们的生成概率模型：

$$p(\theta, z \mid w, \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{1}$$

其中，将 w 作为观察变量，θ 和 z 看作隐藏变量，可以通过 EM 算法学习出 α 和 β。

3.3 句子特征分析

3.3.1 句子权重计算

句子的权重是指句子被选择作为文摘句的依据，权重越大，越有可能是文摘句。我们认为，如果句子的主题与文档要表达的主题是相似的，那么这句话作为摘要的可能性就越大，基于这个思想，在此使用 KL 散度来计算句子主题分布向量与文档主题分布向量之间的离散度：

$$D_{KL}(P \| Q) = \sum_i P(i) * \log \frac{P(i)}{Q(i)} \tag{2}$$

由于 KL 距离是不对称的，即 D_{KL}(P‖Q)与 D_{KL}(Q‖P)是不相同的，所以在实现时取 D_{KL}(P‖Q)与 D_{KL}(Q‖P)的平均值来衡量句子的权重，这样可以使得到的结果更具有代表性。

3.3.2 特征向量建立

运用 LDA 对文档集合建模，可以将文档集合划分成预定的多个主题类，我们将文档表示成 K 维向量空间 V(D)=(wd(T_1), wd(T_2),…, wd(T_n))，其中 T_i 为第 i 个主题，wd(T_i)为给定文档 D 属于主题 T_i 概率 P(T_i|D)，从 θ_d 可以得到每个文档的主题分布 P(T_i|D)。同时，将句子也表示成 K 维向量空间 V(S)=(ws(T_1), ws(T_2),…, ws(T_n))，其中 ws(T_i)为给定句子属于主题 T_i 概率 P(T_i|S)。

在计算句子的主题概率分布时，句子的主题概率分布是以句子中所包含的词语主题概率分布为基础计算的。对于句子 S，其中 S={w_1，w_2，……w_n}，句子的主题分布 P(T_i|S)使用下列公式计算：

$$P(T_i \mid S) = \sum_{W_j \in S} P(T_i \mid W_j) \tag{3}$$

同时，由于在计算句子在主题上的分布的时候，使用了词汇概率的累加，容易使结果偏向长句子，因此对上述公式进行正则化处理，使用词汇概率累加的平均来代替原来的值。所以，句子的主题概率分布公式修改为：

$$P(T_i \mid S) = \frac{\sum_{W_j \in S} P(T_i \mid W_j)}{len(S)} \tag{4}$$

其中，len(S)是句子 S 的长度。

根据贝叶斯定律：

$$P(T_i \mid W_j) * P(W_j) = P(W_j \mid T_i) * P(T_i) \tag{5}$$

我们近似的认为 P(T_i)为 1/K，其中 K 为主题数目，P(W_j)为 1/wordnum，其中 wordnum 为文档中互不相同的单词个数。通过公式(4)、(5)的计算，得到句子属于某个主题的概率，见公式(6)：

$$P(T_i \mid S) = \frac{\sum_{W_j \in S} (P(W_j \mid T_i) * P(T_i)) / P(W_j)}{len(S)} \tag{6}$$

最终得到句子 S 的主题概率分布向量：V(S)=(P(T_1|S), P(T_2|S)…, P(T_n|S))。对于每个句子，我们与文档的主题概率分布向量 V(D)=(wd(T_1), wd(T_2),…, wd(T_n))按照公式(3)计算出句子 S 作为文摘句的权重。

3.4 句子抽取

本系统使用句子抽取的方法为：将多文档集合中所有的句子按照权重统一进行排序，按照排序顺序进行文摘句抽取。根据上面特征计算得到文档集中的各句子的分值，按照分值的大小对句子统一进行排序系统按照句子分值从大到小抽取句子。

通过这种方法可以抽取出一个初步的文摘，但是由于分值的相近，这个文摘抽取出的文摘句的冗余度比较大，需要通过相似度计算来减少文摘中表示同一主题的冗余的句子。我们提出如下计算句子相似度的方法：

$$Sim(S_1, S_2) = \frac{1}{2} * \left(\frac{SameWords(S_1, S_2)}{len(S_1)} + \frac{SameWords(S_1, S_2)}{len(S_2)} \right) \tag{7}$$

式中，S_1，S_2 分别表示两个句子，SameWrods(S_1,S_2)表示 S_1 和 S_2 中相同的词语的个数，len(S_1)和 len(S_2)分别表示句子 S_1 和 S_2 中互不相同的词汇个数。

系统对每个抽取出的句子与已经选择的文摘的每个句子按照公式(8)进行相似度计算，直到抽取出足够的文摘句。采用阈值为 0.7 进行判定，如果相似度的值大于阈值，则认为该主题已经有句子被选择成为文摘句了，例如如下两个句子：

①本月 3 日，一名喝醉的美军士兵于凌晨闯入当地民宅，公然猥亵熟睡中的女中学生。

②本月 3 日凌晨，一名喝醉了酒的冲绳美军普天间机场所属海军陆战队士兵闯入冲绳市内一座公寓，公然对一名熟睡中的女中学生进行猥亵。

对句子①和②用公式(7)计算，系统将会判定这两个句子属于同一个主题从而过滤掉第二句。

4 实验结果与分析

自动文摘的评价按照是否需要人工参与分为人工评价和机器评价，人工评价由于需要人的参与代价非常高，而且由于人的思维的不同使评价中的可变因素增多，本文使用机器评价。机器评价大致分为两种：内部评测和外部评测。内部评测就是测试文摘本身是否与文章的要点一致，以及是否包含文章的基本要点;外部评测就是通过文摘与文章的相似度计算或者文摘在信息检索中所起的作用的大小来评估文摘。

表 1 系统评测结果

文档集合编号	M1			M2		
	Precision	Redundancy	Total	Precision	Redundancy	Total
1	0.58	0.1	0.48	0.9	0.1	0.68
2	0.1	0.1	0	0.56	0.1	0.48
3	0.92	0.1	0.82	1	0.1	0.9
4	0.84	0.1	0.74	0.92	0.1	0.82
5	0.38	0.1	0.28	0.96	0.1	0.9
6	0.48	0.1	0.38	0.68	0.1	0.58
7	0.66	0.1	0.56	0.95	0.1	0.85
8	0.26	0.1	0.16	0.7	0.1	0.6
9	0.48	0.1	0.38	0.36	0.1	0.26
10	0.54	0.1	0.44	0.85	0.1	0.75
11	0.56	0.1	0.46	0.56	0.1	0.46
12	0.48	0.1	0.38	0.9	0.1	0.8
13	0.8	0.1	0.7	0.84	0.1	0.74
14	0.5	0.1	0.4	0.68	0.1	0.68
15	0.38	0.1	0.28	0.66	0.1	0.56
16	0.17	0.1	0.07	0.5	0.1	0.35
17	0.2	0.1	0.1	0.86	0.2	0.76
18	0.44	0.1	0.34	0.77	0.1	0.73
19	0.3	0.1	0.2	0.68	0.1	0.44
平均值	0.477	0.1	0.377	0.754	0.105	0.649

本文采用对每个主题采用模糊标注的方法，标注过程中，除了在源文档集合中标注出标准文摘句，还标注出在源文档中可替换标准文摘句、且不能与标准文摘句在文摘中同现的句子，我们称之为候选文摘句。每个候选文摘句根据可替换程度赋予一个取值在(0，1]之间的权值。这样得到的评测语料库就可以采用准确率、冗余度和总体质量三项指标来评估文摘系统质量，以解决传统多文档自动文摘评测出现的无法顾全文本集合中存在多个可替换文摘句的问题。在此基础上，采用准确率、冗余度和综合质量等几方面指标来评估待测系统：

$$\Pr ecision = (\sum_{i=1}^{K} \varphi_i) / K$$

$$\mathrm{Re}\, dundancy = (\sum_{i=1}^{K} (\sum_{j=i+1}^{K} \varphi(S_i, S_j))) / K$$

$$Total(summary) = \Pr ecision - Redundancy$$

本文的实验语料使用哈工大多文档自动文摘评测平台提供的原始语料，该语料包含 19 个主题 117 篇新闻文档，每组包含 5~10 篇文档。三项指标都是按照 10 句文摘长度测试的，M1 是传统文摘算法，采用句子中词汇的 TFIDF 特征加权平均作为句子的权重。M2 是 LDA 建模的主题分布算法。表 1 为实验数据。

表 1 数据表明，根据主题分布计算句子重要度的方法得到的文摘结果，准确度要优于传统的文摘算法得到的文摘结果。从表 1 中可以看出，19 个文档集合分别在两种方法下文摘的准确率按同一趋势增长，但冗余度变化不大，由此说明这种区分句子冗余度的方法还需要改进。

5 总结

本文提出了一种基于 LDA（Latent Dirichlet Allocation）主题概率分布模型的中文多文档自动文摘方法。该方法使用 LDA 为多文档集合建模，得到句子的主题概率分布和文档的主题概率分布，通过计算概率分布之间的相似度作为句子的重要度，并根据句子重要度进行文摘句的抽取。实验表明，该方法所得到的文摘，性能优于传统文摘方法。但由于该方法仍然倾向于抽取长句，而长句子影响了系统的性能，以后的工作将考虑在性能提高的基础上对抽取出的文摘进行压缩。另外，由于 LDA 模型中使用词袋假设，因此不能很好的表示句子、文档和文档集合之间的结构关系，我们将在今后的工作中扩充 LDA 模型，使之更加适用于自动文摘。

参考文献

[1]. RADEV D R, MCKEOVWN K R. Generating natural languages summaries from multiple on-line sources[J]. Computational Linguistics, 1998, 24(3): 21-29.

[2]. LIN C Y, HOVY E. From single to multi-document summarization: a prototype system and its evaluation[C]. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia: ACL, 2002. 457-464.

[3]. 王萌，李春贵，唐培和等. 一种主题句发现的中文自动文摘研究[J]. 计算机工程，2007，33(8): 180-181.

[4]. 秦兵，刘挺，李生. 多文档自动文摘综述[J]. 中文信息学报，2005，19(6): 13-20.

[5]. Rachit Arora. Latent Dirichlet Allocation Based Multi-Document Summarization[C]. Proceedings of the second workshop on Analytics for noisy unstructured text data. Pages: 91-97,2008.

[6]. Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis[C]. In Proceedings of ACM SIGIR 2001. Pages19-25, Louisiana, USA. Sep. 2001.

[7]. Harendra Bhandari, Masashi Shimbo, Takahiko Ito, and Yuji Matsumoto. Generic text summarization using probabilistic latent semantic indexing [C]. Proceedings of IJCNLP 2008, Pages 133-140.

[8]. 吴晓峰，宗成庆. 一种基于 LDA 的 CRF 自动文摘方法[J]. 中文信息学报，2009，23(6): 39-45.

[9]. David.M.Blei, Andrew Y.Ng and Michael I. Jordan. Latent Dirichlet Allocation [J]. Journal of machine Learning Research 3. Pages: 993-1022, 2003.

[10]. Aria Haghighi, Exploring Content Models for Multi-Document Summarization[C]. The 2009 Annual Conference of the North American Chapter of the ACL, pages: 362-370.

[11]. 刘茂福，李淑君，金可佳等. 多文档自动文摘中的特征组合优化[J]. 计算机系统应用， 2008，17(8): 59-63.