Computer Science and Application 计算机科学与应用, 2013, 3, 145-148
http://dx.doi.org/10.12677/csa.2013.32025 Published Online April 2013 (http://www.hanspub.org/journal/csa.html)

Automatic Summarization Algorithm Based on the Combined Features of LDA*

Dengneng Wu, Zhenming Yuan, Xingxing Li

Hangzhou Normal University, Hangzhou
Email: wudn@stu.hznu.edu.cn, zmyuan@hznu.edu.cn, xxl@stu.hznu.edu.cn

Received: Feb. 25th, 2013; revised: Mar. 12th, 2013; accepted: Mar. 21st, 2013

Abstract: Automatic summarization can help people to get the main information from the massive amounts of information more quickly and efficiently. In this paper, a document summarization algorithm based on LDA is proposed. Firstly, we calculate the similarity of topics probability distribution between document and sentence as a new feature. Then, we also considered traditional summarization features such as position of sentence in a text and topic similarity. Finally, summary are generated by selecting the sentences with highest scores. Experimental results show that the performance of our method outperforms the traditional methods when the combined features join into the LDA Model.

Keywords: Automatic Summarization; Topic Model; LDA

基于组合特征 LDA 的文档自动摘要算法*

吴登能，袁贞明，李星星

杭州师范大学，杭州
Email: wudn@stu.hznu.edu.cn, zmyuan@hznu.edu.cn, xxl@stu.hznu.edu.cn

收稿日期：2013 年 2 月 25 日；修回日期：2013 年 3 月 12 日；录用日期：2013 年 3 月 21 日

摘　要：文档自动摘要可以帮助人们在海量信息中快速高效地获取主要信息。本文以句子作为处理单元，提出一个基于 LDA 模型的句子主题特征，通过计算文档主题分布与句子主题分布之间的相似性，结合句子在文档中的位置和标题相似性等基础特征，形成组合特征计算句子权重，最后根据权重排序抽取摘要。实验结果显示，在 LDA 模型中加入组合特征后，自动摘要的性能得到了提高。

关键词：自动摘要；主题模型；LDA

1. 引言

随着移动互联网的广泛应用，信息数量激增，用户面临信息过载问题。受到移动终端的屏幕大小和连接带宽大小等限制，推送给用户的新闻通常首先要做摘要处理。按照摘要产生方法的不同自动摘要可以分为抽取式摘要(extractive)和理解式摘要(abstractive)[1]。抽取式摘要直接从原文中抽取重要的句子作为摘要句，而理解式摘要则通过对文章进行句法、语义和篇章结构的分析获取文档的意义，再通过自然语言生成得到满足要求的摘要[2,3]。

基于 LDA[4] (Latent Dirichlet Allocation)的抽取式摘要是近期的研究热点。Shafiei[5]提出一种由词、片段、主题、文档四层结构组成的 Co-Clustering Model 模型，该方法受限于摘要长度，并不是所有从主题类

*资助信息：浙江省自然科学基金重点项目(Z12F020027)；教育部 211 重点工程项目(201003017)。

中选出的句子都能作为摘要内容，使得产生的摘要内容代表性不强。Haghighi[6]将句子、文档和文档集合统一纳入到一个层次性 LDA 主题模型中，使用 Gibbs 抽样获得模型参数，以 KL-散度作为摘要评价模型选择句子，使用贪心算法添加句子。Arora 等[7]使用 LDA 作为文档的表示模型，提出了基于推论的、半生成性和全生成性的 3 种句子选择形式。该方法仅仅通过计算句子的主题概率来选择摘要句子，忽略了其他常用特征，使得选出的摘要质量不高。

鉴于以上几种方法的缺点，本文提出一种基于 LDA 模型的句子主题特征，以句子作为处理单元，根据 LDA 模型中主题的概率分布和句子的概率分布计算文档与句子的主题相似性，并融合句子在文档中的位置和标题相似性等基础特征，形成组合特征共同评价句子的重要性，最后根据融合特征分值大小抽取句子生成摘要。本文系统流程如图 1 所示。

2. 基于组合特征 LDA 的自动摘要算法

本文首先提取句子的基础特征，然后分别提取文档和句子的 LDA 主题概率分布模型，得到句子模型和文档模型之间的主题相似性特征，最后融合以上两类特征抽取得分高的句子作为摘要。

2.1. 句子的基础特征

基础特征表征句子在文档中的重要程度，包括：句子长度、位置和标题相似度等特征[8]。

1) 长度特征 $L(S)$：避免摘要偏向于长句子，为句子的长度添加权重值，定义句子 S 的长度特征为：

$$L(S)=1-\frac{|t-\mu|}{\mu} \tag{1}$$

其中，μ 为同一文档集合中句子的平均长度，t 为句子的长度。

2) 位置特征 $Pos(S)$：假设文章有 N 个句子，S 为其中的第 k 个句子，定义句子 S 的位置特征为：

$$Pos(S)=\frac{N-k+1}{N} \tag{2}$$

3) 相似度特征 $Sim(S,T)$：把一篇文档的标题看成最主要的句子，将标题句与文档中每个句子向量化，计算它们之间的相似度，相似度较大者该特征项得分较高。

$$Sim(S,T)=\frac{S\cdot T}{|S||T|} \tag{3}$$

本文将按照这四个基本特征抽取的摘要系统作为 Baseline，Baseline 系统中句子 S 的基础特征为上面四个特征值的加权和：

$$\text{ScoreBase}=L(S)+2\times Pos(S)+Sim(S,T) \tag{4}$$

对于新闻文档而言，其中句子的位置对该句子能否作为摘要句有更好的指示，因而相对权重较大。

2.2. 句子的 LDA 主题概率特征

为表示词汇、句子和文档之间的关系，本文采用四层 LDA 模型[9]，分别提取文档和句子的 LDA 主题概率分布模型，得到计算句子模型和文档模型之间的主题相似性特征。

2.2.1. 句子权重计算

一般而言，如果一篇文档中句子的主题与文档所要表达的主题越相似，那么这句话作为摘要句子的概率就越大。基于这个思想，通过计算句子的主题概率分布与文档主题概率分布的相似度来评价句子的重要度，这里使用 KL 散度来计算两个分布之间的相似度。

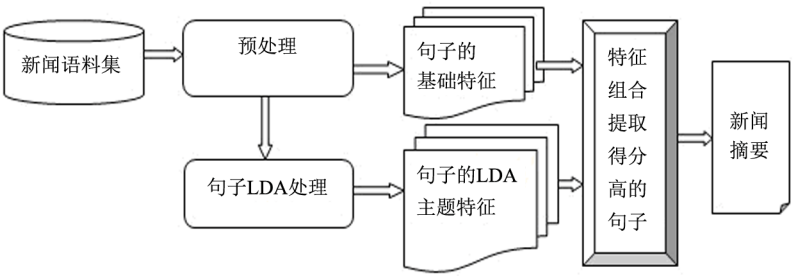

Figure 1. Overview of our algorithm
图 1. 基于组合特征 LDA 的摘要算法流程

$$TSim(S,D) = -\left(D_{KL}\left(P_S \| P_D\right) + D_{KL}\left(P_D \| P_S\right)\right) \quad (5)$$

式中，P_S 和 P_D 分别表示句子 S 和文档 D 在主题上的概率分布，$D_{KL}\left(P_S \| P_D\right)$ 是两个概率分布 P_S 和 P_D 之间的 KL 散度，即：

$$D_{KL}\left(P_S \| P_D\right) = \sum_i P_S(i) \log \frac{P_S(i)}{P_D(i)} \quad (6)$$

2.2.2. 主题概率分布计算

LDA 模型是一个描述如何基于潜在主题生成文档中词的概率抽样过程，其生成过程如下[9]：

1) 对文集中的任一文档 d，生成一个 θ~Dir(a)；

2) 考虑文档 d 中的词 w_i 的生成：

a) 生成一个主题 z_j ~ Multinomial(θ)；

b) 对话题 z_j 生成一个离散变量 φ_z ~ Dir(β)；

c) 生成使得 $P\left(w_i \middle| \varphi_z, \beta\right)$ 最大的一个词。

其中，a 的值表示各个主题在取样之前的权重分布，β 的值表示各个主题对词的先验分布。

运用 LDA 对文档集合建模，将新闻文档划分成预定的 K 个主题，因此文档可表示成 K 维的向量空间 $P_D = \left\{P\left(T_1 \middle| D\right), P\left(T_2 \middle| D\right), \cdots, P\left(T_k \middle| D\right)\right\}$，其中 T_i 为第 i 个主题，分量 $P\left(T_i \middle| D\right)$ 为给定文档 D 属于主题 T_i 的概率，在 LDA 模型中可以从 θ_d 得到每个文档的主题分布 $P\left(T_i \middle| D\right)$。句子也被表示成 K 维向量空间 $P_S = \left\{P\left(T_1 \middle| S\right), P\left(T_2 \middle| S\right), \cdots, P\left(T_k \middle| S\right)\right\}$，其中每个分量 $P\left(T_i \middle| S\right)$ 为给定句子 S 属于主题 T_i 的概率。

句子是由词汇组成，因此句子的主题概率分布可以由句子中所包含的词汇主题概率分布计算得到。对于句子 S，其中 $S = \left\{W_1, W_2, \cdots, W_n\right\}$，$W_i$ 表示句子中的单词，则句子的主题分布 $P\left(T_i \middle| S\right)$ 可由下列公式计算得到：

$$P\left(T_i \middle| S\right) = \sum_{W_j \in S} P\left(W_j \middle| T_i\right) \times P\left(T_i \middle| D\right) \times P(D) \quad (7)$$

为了计算简单，假设文档集中的文档初始是等概率的，则句子的主题分布 $P\left(T_i \middle| S\right)$ 可简化为下列公式：

$$P\left(T_i \middle| S\right) = \sum_{W_j \in S} P\left(W_j \middle| T_i\right) \times P\left(T_i \middle| D\right) \quad (8)$$

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。句子的主题概率分布公式修改为：

$$P\left(T_i \middle| S\right) = \frac{\sum_{W_j \in S} P\left(W_j \middle| T_i\right) \times P\left(T_i \middle| D\right)}{\text{len}(S)} \quad (9)$$

其中 $P\left(W_j \middle| T_i\right)$ 是 LDA 的输出结果，因此得到句子 S 的主题概率分布向量 P_S。文档 D 的主题概率分布向量 P_D，则由 LDA 的输出结果 $P\left(T_i \middle| D\right)$ 作为它的分量。

2.3. 融合组合特征的关键句抽取

根据上面组合特征计算得到新闻文档中的各句子的分值，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，直到达到摘要指定的数量。可由下列伪代码表示：

Algorithm Sentences weight calculation

Input:

*Given two sets of the candidate sentences and the summary sentences $O=\{o_1,o_2,...o_n\}$ and $S=\{s_1,s_2,\ldots s_m\}$

Output:

*The scores of the sentence set S

Procedure

1: For sentence o_i in O, do;

Using Eq(4);

Obtain ScoreBase(o_i);

Using Eq(5) and Eq(9);

Obtain scores(o_i) =TSim(o_i,d);

2: End for

3: Select m sentences from O to compose set S, which has the first m maximum values of scores(O)+ScoreBase(O).

3. 实验结果

本文采用 DUC2007 自动摘要评测任务中的数据集和网络爬虫从网易网站抓取新闻报道数据作为实验数据，DUC2007 包含 45 个文档集合，每个集合包含 25 篇来具有共同话题或相关话题的文档，使用软件对文档句子进行划分。网络抓取数据集覆盖的主题有体育、经济、事故等，整个数据分成 15 个文档集合，每个文档集合包含 10 篇文档，相同文档集合中的文档拥有一个共同中心主题。

实验为每个文档集合建立 LDA 模型，根据本文提出特征算法抽取出长度为 200 个单词的自动摘要并使用评测工具 ROUGE 自动评测摘要结果。采用 ROUGE-1，ROUGE-2，ROUGE-3，ROUGE-L 这四

个评测标准。ROUGE-n 代表的是基于 n-unigram 的共现统计，它反映机器摘要与理想摘要中共现的单词在理想摘要中的比率。

实验结果分析

为了体现我们算法的效果，我们与 Baseline，Doc-LDA[3]，KL-LDA[7]，LSA[6]等几种摘要算法做了对比。

- Baseline：基于词频统计等传统特征的文档摘要方法。
- KL-LDA：将句子、文档和文档集合统一纳入到主题模型中，以 KL-散度作为摘要评价模型选择句子，使用贪心算法添加句子。
- LSA：对句子进行奇异值分解，按得分高低选择句子作为摘要句。
- Doc-LDA：根据主题概率大小排序，然后从大到小选择主题，再从主题中选择概率大的句子作为摘要句。

表 1 给出在数据集 DUC2007 上各模型的实验结果。表 2 描述各模型在网络数据集上的实验效果。可以看出，根据 ROUGE 的 4 个评测标准判断的各模型性能的好坏相差不大。用奇异值分解的模型 LSA 效果最不理想，其他 3 种模型的效果相当，本文提出算法效果总体上优于其他几种对比算法效果。

Table 1. The experimental evaluation results of DUC2007
表 1. DUC2007 数据实验评测结果

Algorithm	Rough-1	Rough-2	Rough-3	Rough-L
Baseline	0.32018	0.05822	0.01214	0.29113
KL-LDA	0.31346	0.06141	0.01821	0.26548
LSA	0.25847	0.03631	0.00857	0.22671
Doc-LDA	0.30251	0.05272	0.01091	0.26632
Our-LDA	0.34722	0.06412	0.01912	0.30368

Table 2. The experimental evaluation results of web data
表 2. 网络数据集实验评测结果

Algorithm	Rough-1	Rough-2	Rough-3	Rough-L
Baseline	0.28506	0.04259	0.01038	0.25896
KL-LDA	0.27712	0.04698	0.01173	0.23234
LSA	0.25749	0.03682	0.00861	0.23294
Doc-LDA	0.29937	0.05159	0.01015	0.25832
Our-LDA	0.33452	0.05926	0.01392	0.29534

4. 总结

本文提出了一种基于 LDA 主题概率分布模型的新闻文档自动摘要方法。方法使用 LDA 为新闻文档集合建模，得到句子的主题概率分布和文档的主题概率分布，通过计算两者概率分布之间的相似度作为句子的重要度，再组合传统基于词频等特征，对新闻文档的句子进行打分，句子按得分高低重新排列，最后抽取句子生成所需摘要。实验结果表明，该方法的性能优于传统摘要方法。

参考文献 (References)

[1] Y. D. Xu, Z. M. Xu and X. L. Wang. Multi-document automatic summarization technique based on information fusion. Chinese Journal of Computers, 2007, 30(11): 2048-2054.

[2] W. Gao, P. Li and K. Darwish. Joint topic modeling for event summarization across news and social media streams. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, New York: ACM, 2012: 1173-1182.

[3] A. Celikyilmaz, D. Hakkani-Tur. A hybrid hierarchical model for multi-document summarization. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Stroudsburg: Association for Computational Linguistics, 2010: 815-824.

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

[5] M. M. Shafiei, E. E. Milios. Latent Dirichlet co-clustering. Proceedings of the 6th International Conference on Data Mining (ICDM), Hong Kong, 18-22 December 2006: 542-551.

[6] A. Haghighi, L. Vanderwende. Exploring content models for multi document summarization. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Colorado: Association for Computational Linguistics 2009: 362-370.

[7] R. Arora, B. Ravindran. Latent Dirichlet allocation based multi-document summarization. Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data, Singapore, 2008: 91-97.

[8] L. Maofu, L. Shujun and J. Kejia. Combination optimization of features in multi-documents automatic summarization. Computer Systems & Applications, 2008, 17(8): 59-63.

[9] Y. L. Chang, J. T. Chien. Latent Dirichlet learning for document summarization. IEEE International Conference on Acoustics, Speech, and Signal Processing, Taipei, 19-24 April 2009: 1689-1692.