基于词语对狄利克雷过程的时序摘要

席耀一 [1] 李弼程 [1] 李天彩 [1] 黄山奇 [2]

摘 要 时序摘要是按照时间顺序生成摘要, 对话题的演化发展进行概括. 已有的相关研究忽视或者不能准确发现句子中隐含的子话题信息. 针对该问题, 本文建立了一种新的主题模型, 即词语对狄利克雷过程, 并提出了一种基于该模型的时序摘要生成方法. 首先通过模型推理得到句子的子话题分布; 然后利用该分布计算句子的相关度和新颖度; 最后按时间顺序抽取与话题相关且新颖度高的句子组成时序摘要. 实验结果表明, 本文方法较目前的代表性研究方法生成了更高质量的时序摘要.

关键词 时序摘要, 狄利克雷过程, 词语对, 主题模型

引用格式 席耀一, 李弼程, 李天彩, 黄山奇. 基于词语对狄利克雷过程的时序摘要. 自动化学报, 2015, **41**(8): 1452−1460

DOI 10.16383/j.aas.2015.c150001

Temporal Summarization Based on Biterm Dirichlet Process

XI Yao-Yi[1] LI Bi-Cheng[1] LI Tian-Cai[1] HUANG Shan-Qi[2]

Abstract Temporal summarization aims at extracting sentences chronologically to give an overview about the evolution of a topic. Existing researches either neglect the information of latent subtopics, or fail to accurately discover them. In this paper, we develop a novel topic model called biterm Dirichlet process and generate the temporal summary based on it. Firstly, we get the subtopic distribution in each sentence through posterior inference. Secondly, we calculate each sentence′s relevance and novelty degree according to its subtopic distribution. Finally, we chronologically extract the sentences which are relevant and novel to generate the temporal summary. Experiments demonstrate the better performance of our approach compared with currently representative methods.

Key words Temporal summarization, Dirichlet process, biterm, topic model

Citation Xi Yao-Yi, Li Bi-Cheng, Li Tian-Cai, Huang Shan-Qi. Temporal summarization based on biterm Dirichlet process. *Acta Automatica Sinica*, 2015, **41**(8): 1452−1460

时序摘要 (Temporal summarization) 是传统多文档摘要与更新摘要技术的延伸与发展, 指的是对于某一话题, 按照其发展的时间顺序对不同阶段的相关文档进行摘要分析, 要求每一时间段的摘要不能包含之前时间段内已出现的话题信息, 即时序摘要要求随着时间的发展, 对摘要进行动态更新, 不断添加随话题发展新出现的信息, 因此又称之为序列更新摘要 (Sequential update summarization)、时间线摘要 (Timeline summarization) 或者故事线摘要 (Storyline summarization).

目前, 国内外很多学者都针对时序摘要开展了相关研究. 已有的方法均是通过抽取原文句子并组合得到摘要, 主要分为以下三类: 基于时间分段的方法、基于句子簇的方法和基于子话题发现的方法. 基于时间分段的方法[1−3] 一般按照文档的发布时间将文档集分段, 然后分别生成每一时间段的摘要, 并将各个时间段的摘要按照时间顺序拼接得到时序摘要. 该方法简单直观, 但是由于新闻报道的重复性特点, 使得很多内容相同的句子并没有随时间分段而分开, 增加了冗余信息被误判为摘要句的可能性; 基于句子簇的方法[4−5] 认为如果两个句子所报道事件的发生时间比较接近, 那么这两个句子属于同一子话题的可能性较大, 该方法需要对句子进行时间消解, 难度较大; 基于子话题发现的方法能够挖掘话题的语义信息, 便于摘要句的选择.

话题一般由若干子话题组成, 例如 "9·11 恐怖袭击" 话题通常包含以下几个子话题: 恐怖袭击发生、美国政府紧急应对、国际社会强烈谴责等. 文档集中的句子一般与某一子话题或几个子话题相关, 而基于时间分段的方法和基于句子簇的方法通常忽视了该相关性, 没有利用深层语义信息, 仅是通过浅层次的词语特征信息选择句子. 主题模型自 2003 年由 Blei 等[6] 正式提出以来, 就被应用于挖掘文档集中的隐含主题[7], 取得了不错的效果. 在时序摘要研究中, 已有学者尝试利用主题模型发现隐含子话题,

收稿日期 2015-01-04 录用日期 2015-04-08
Manuscript received January 4, 2015; accepted April 8, 2015
国家社会科学基金 (14BXW028) 资助
Supported by National Social Science Foundation of China (14BXW028)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 解放军信息工程大学信息系统工程学院 郑州 450001 2. 65022 部队 沈阳 110162
1. Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou 450001 2. Unit 65022, Shenyang 110162

并辅助生成摘要. Gao 等[8] 将动态隐含狄利克雷分配模型应用于 Twitter 数据流的摘要生成当中. Huang 等[9] 提出 Mixture-event-aspect model, 用于发现全局子话题和局部子话题, 并根据子话题出现的时间顺序分别生成相应的摘要. Li 等[10] 提出演化分层狄利克雷过程对话题演化进行建模, 进而生成时序摘要, 取得了比较好的性能.

为了抽取摘要句, 文献 [8−10] 均尝试利用主题模型发现每一 Twitter 或者句子的隐含子话题. 由于主题模型依赖于词语共现挖掘隐含子话题, 而 Twitter 和句子均属于短文本, 特征稀疏, 词语共现信息少, 不利于传统主题模型的应用. 为解决这一问题, Yan 等[11] 提出了词语对主题模型 (Biterm topic model, BTM)[1]. 该模型跳出单一短文本的限制, 直接从文档集中提取所有特征的共现信息, 避免受到单一短文本特征稀疏的影响. BTM 模型假设文档集中子话题的先验分布服从狄利克雷分布, 导致其必须事先指定主题个数. 然而, 实际应用中面临的数据对用户来讲一般是未知的, 不可能准确知道其中隐含的主题数目, 这降低了 BTM 模型的应用价值. 为解决上述问题, 本文建立了一种新的主题模型, 即词语对狄利克雷过程 (Biterm Dirichlet process, BDP), 并利用该模型抽取句子生成时序摘要. BDP 将单个句子视为一篇文档, 以狄利克雷过程作为文档集中子话题分布的先验分布, 可以自动确定文档集的子话题数目. 首先, 通过模型推理得到每一句子的子话题分布; 然后, 根据时序摘要的特点, 利用该分布计算句子的相关度和新颖度; 最后, 按时间顺序抽取与话题相关且新颖度高的句子组成时序摘要.

本文的组织结构如下: 第 1 节介绍了相关研究现状, 第 2 节详细介绍了本文生成时序摘要的方法, 最后给出了实验结果与分析.

1 相关工作

关于时序摘要的最早研究可以追溯至 2001 年, Allan 等[12] 首次提出时序摘要的概念, 然而由于当时的文档摘要技术还不成熟且时序摘要研究面临的困难较多, 之后几年研究关注的焦点仍然是传统的多文档摘要技术. 2004 年, Chieu 等[4] 首先解析每一句子所描述事件的时间, 然后根据每一句子的兴趣度和突发性对句子进行排序, 经过去冗余后根据时间顺序生成时间线摘要. 2008 年, Lin 等[13] 研究了基于故事线的摘要来分析话题, 首先识别事件, 然后构建主故事线, 最后为主故事线中的每一事件生成摘要, 并按照事件顺序前后相连得到基于故事线的摘要. 然而, 将每一事件的摘要直接相连不可避免会导致最终的摘要中存在冗余. 2009 年, 贺瑞芳等[14] 提出了基于宏微观重要性判别模型的时序多文档摘要方法. 该方法通过宏观重要性判别模型判断不同时间段的重要性, 通过微观重要性判别模型判断句子的重要性并选择重要程度高的句子组成每一时间段的摘要, 最后将重要时间段的摘要按照时间顺序前后连接得到最终的时序摘要. 贺瑞芳等的研究虽然考虑了对不同时间段的重要性进行区分, 但是假设冗余信息仅存在于各个时间段内部, 与实际不符. 2011 年, Yan 等[1−2] 提出进行时间线摘要的研究, 尝试以摘要的形式给出话题的演化轨迹. 其方法的核心是构造目标函数, 然后通过全局和局部最优化选择句子生成摘要. Yan 等的研究既考虑了区分不同时间段的重要性, 又考虑了通过最优化的方式解决摘要信息的冗余问题, 然而其还没有考虑利用话题中的隐含语义信息. 2012 年, Chen 等[15] 将矩阵奇异值分解引入时序摘要的生成当中, 首先检测 themes; 然后识别每一 theme 中的事件, 同时生成每一事件的摘要; 最后识别事件之间的演化关系. Chen 等的方法仍然没有考虑滤除摘要内的冗余信息. 2013 年, Huang 等[9] 在生成摘要的过程中考虑利用全局子话题和局部子话题信息, 但是需要事先确定子话题数目. Xu 等[5] 首次考虑将不同的子话题分别以时间线摘要的形式展示给用户, 提出了二维故事图的概念, 即以图展示整个话题的时序摘要, 图中包含多条故事线, 每一故事线均对应了一个子话题. Li 等[10] 研究了基于演化分层狄利克雷过程的时间线摘要生成方法, 其将文档按时间段划分之后, 利用一个三层狄利克雷过程进行建模, 既能识别每一时间段的主题, 又能获取不同时间段主题之间的关系.

上述研究均是在话题相关文档已知的前提下进行的, 属于静态时序摘要的范畴. 为促进动态时序摘要的研究, 美国国家标准与技术研究院 (National Institute of Standards and Technology, NIST) 于 2012 年提出进行时序摘要的国际评测, 并将其作为 2013 年 TREC 国际评测会议 (Text retrieval conference) 的任务之一[16]. 该评测吸引了国内外多家研究机构的参与, 但是从评测结果来看, 动态时序摘要还有待更深入的研究.

另外, 近年来随着社交媒体的迅速发展, 研究者尝试对社交媒体数据流进行摘要分析, 以帮助不同需求的用户能够从大规模数据流中快速掌握感兴趣话题的主要信息. 这一研究又称为数据流摘要[17−20]. 与时序摘要主要考虑对单一话题进行摘要分析不同, 数据流摘要面向的是包含多个话题的社

[1]Biterm 在本文中指的是共现词语对. 例如句子 “中国战胜日本” 中包含三个共现词语对, 分别是 “中国 – 战胜”、“中国 – 日本” 和 “战胜 – 日本”.

交媒体数据流, 要求从不同时期的数据流中抽取代表性的 Twitter 博文作为其摘要.

2 研究方法

基于 BDP 的时序摘要生成方法首先建立 BDP 模型发现文档集中的隐含子话题, 并得到每一句子的子话题分布; 然后, 利用该分布计算句子的相关度和新颖度, 选择既相关又新颖的句子; 最后, 对选择的句子按照时间顺序排序组成时序摘要.

2.1 子话题发现

本文利用 BDP 发现文档集中隐含的子话题. BDP 在建模过程中, 打破了单一文档的边界限制, 直接根据文档集内的词语共现信息挖掘隐含子话题.

2.1.1 BDP 模型建立

BDP 模型假设文档集中的主题先验分布服从狄利克雷过程, 如图 1 所示.

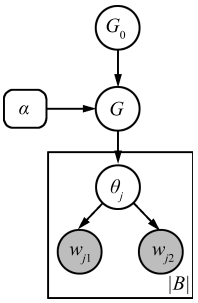


图 1 词语对狄利克雷过程

Fig. 1 Biterm Dirichlet process

狄利克雷过程是基于分布上的分布. 对于测度空间 $(\Theta, \mathbb{B})$ 上的概率分布 G 和 G_0, 如果对 $(\Theta, \mathbb{B})$ 的任意有限分割 $(T_1, T_2, \cdots, T_K)$, G 均满足

$$(G(T_1), G(T_2), \cdots, G(T_K)) \sim Dir(\alpha G_0(T_1), \alpha G_0(T_2), \cdots, \alpha G_0(T_K)) \quad (1)$$

则称 G 满足狄利克雷过程, 记为 $G \sim DP(\alpha, G_0)$. 其中, α 是聚焦参数, 为正实数 (Positive real number), G_0 为基分布.

目前, 常采用的狄利克雷过程构造方法有三种, 分别是截棍过程 (Stick-breaking construction)、中国餐馆过程 (Chinese restaurant process, CRP) 和波利亚罐子模型 (Pólya urn scheme). 利用 CRP 构造狄利克雷过程时, 将文档集视为一个中国餐馆, 每个顾客视为一个 Biterm. 假设一个中国餐馆能容纳无数张桌子, 每张桌子能容纳下无数顾客. 进入餐馆中的每个顾客用 i 表示. 第一个顾客 1 进入餐馆后选定第一张桌子坐下; 第二个顾客 2 进入餐馆后要么会和第一个顾客坐在一起, 要么会另选一张桌子坐下, $\cdots$. 那么, 第 n 个顾客会以 $n_j/(n-1+\alpha)$ 的概率选择已经占有的桌子 θ_j, 其中 n_j 表示选择该桌子坐下的顾客数; 以 $\alpha/(n-1+\alpha)$ 的概率选择一张没有顾客坐的新桌子. 顾客被分到已经有顾客坐的桌子的概率主要和 n_j 相关, 这样导致 "一个坐的人越多的桌子, 再有人坐的可能性越大"; 而顾客坐到没有人坐的桌子的概率主要和 α 相关.

狄利克雷过程认为一组数据是由一个混合模型 (Mixture model) 抽样产生的. 该模型包含多个混合成分 (Mixture components), 每个数据都与一个混合成分相关. 如果将 CRP 中的桌子视为隐含子话题, 那么利用 BDP 生成文档集的过程如下:

1) 选取基分布 $G \sim DP(\alpha, G_0)$, 表示文档集内的子话题分布;

2) 对于文档集内的每一个 Biterm, 将其记为 b_j, 从 G 中抽取子话题 θ_j, 然后根据子话题 θ_j 生成 b_j 所包含的两个词语 w_{j1} 和 w_{j2}.

2.1.2 模型推理

根据 CRP 构造 BDP 模型的过程, 本文采用 Gibbs 采样算法对 BDP 模型进行近似后验推理. 由于篇幅有限, 本文仅给出 Gibbs 采样的迭代式, 详细推导可参考文献 [21−22]. 每一符号的具体含义详见表 1.

表 1 BDP 推理中部分符号的含义

Table 1 Notations in BDP inference

符号	含义
m_z	由子话题 z 生成的 Biterm 数
n_z^v	由子话题 z 生成的特征 v 的数目
V	特征数目, 即词典大小
$\|B\|$	Biterm 总数

第 j 个顾客选择桌子 z 坐下的条件概率分布为

$$p(\theta_j = z | \boldsymbol{z}^{-j}, \boldsymbol{B}^{-j}) \propto \begin{cases} m_z^{-j} f_z^{-w_{j1}}(w_{j1}) f_z^{-w_{j2}}(w_{j2}), & z \text{ has been used} \\ \alpha f_{z^{new}}^{-w_{j1}}(w_{j1}) f_{z^{new}}^{-w_{j2}}(w_{j2}), & z = z^{\text{new}} \end{cases} \quad (2)$$

本文假设每一子话题均服从对称狄利克雷分布 $\varphi_z \sim Dir(\beta)$, 则

$$f_z^{-w_{ji}}(w_{ji} = v) = \begin{cases} \dfrac{n_z{}^{-w_{ji},v} + \beta}{n_z{}^{-w_{ji}} + V\beta}, & z \text{ has been used} \\ \dfrac{1}{V}, & z = z^{\text{new}} \end{cases}, \quad j = 1, \cdots, |B| \quad (3)$$

上标中的 "$^{-}$" 表示不包含相应变量的计数, 例如 m_z^{-j} 表示由子话题 z 生成的 Biterm 数 (不包括第 j 个 Biterm). 通过模型推理可以得到文档集中的子话题数目 K、每一子话题的词语概率分布 f_z^w (又记为 $p(w|z)$)、文档集内每一子话题的概率分布 $p(z)$ 以及每一句子 s 的子话题概率分布 $p(z|s)$, 具体计算方法如下:

$$f_z^w = \frac{{n_z}^w + \beta}{\sum\limits_w {n_z}^w + V\beta} \tag{4}$$

$$p(z) = \frac{n_z + \alpha}{|B| + K\alpha} \tag{5}$$

$$p(z|s) = \sum_b p(z|b)p(b|s) \tag{6}$$

其中, $p(z|b)$ 和 $p(b|s)$ 采用文献 [11] 中的方法计算得到, 如式 (7) 和式 (8) 所示, 即根据贝叶斯公式计算 $p(z|b)$, 采用每一 Biterm 在句子 s 中的经典分布估计 $p(b|s)$.

$$p(z|b) = \frac{p(z)p(w_i|z)p(w_j|z)}{\sum\limits_{z=1}^{K} p(z)p(w_i|z)p(w_j|z)} \tag{7}$$

$$p(b|s) = \frac{m_s(b)}{\sum\limits_b m_s(b)} \tag{8}$$

2.2　摘要句子选择

完成子话题发现后, 如何从文档集合中选择既重要又新颖的句子是一个关键. 时序摘要相比多文档摘要考虑了时间因素. 如果将多文档摘要视为空间中的一个点, 那么时序摘要可以视为由多个点在时间轴上连成的一条线. 多文档摘要需要满足: 1) 覆盖率高, 即摘要内容尽可能覆盖文档集的主要信息; 2) 冗余信息少, 即摘要内的重复信息应尽可能少. 这两个要求在考虑时间维度的前提下又多了一层新的含义. 首先是覆盖率高, 由于话题随时间在不断演化, 因此每一时间段的话题内容之间一般都会有所区别, 时序摘要需要覆盖每一时间段的主要内容; 其次是新颖信息多, 冗余信息少, 时序摘要既要求同一时间段内的冗余信息尽可能少, 又要保证不同时间段之间的冗余信息尽可能少, 即需要进行跨时间段冗余信息滤除 (Cross-date redundancy removal).

时序摘要应该既能反映出话题内容的变化, 又尽量避免冗余信息. 每一时间段的摘要应重点体现该时间段内话题的主要内容, 同时避免与之前时间段的话题内容出现重复. 根据时序摘要的特点, 在每一时间段, 选择那些与当前时间段话题主要内容密切相关, 且包含没有在之前时间段出现过的话题新信息的句子添加入摘要. 本文根据句子的相关度和新颖度选择句子, 其分别定义如下.

句子相关度. 句子与当前时间段文档集所描述话题信息的相关程度.

句子新颖度. 句子与当前时间段之前文档集包含信息的重叠程度.

对于时间段 T 的句子集合 S^T, 以每一句子与该时间段的主题分布的相似性作为其相关度的度量, 相似度越高则相关性越强. 引入递减 logistic 函数 $\zeta_1(x) = 1/(1+\mathrm{e}^x)$ 和 KL 散度 (Kullback-Leibler divergence), 每一句子的相关度计算如下:

$$F_{IR}(s^T) = \zeta_1(KL(s^T \| TopicDis^T)) \tag{9}$$

每一句子的新颖度则以其与时刻 T 之前的文档集的主题分布的距离进行度量, 距离越远则新颖度越大. 引入递增 logistic 函数 $\zeta_2(x) = \mathrm{e}^x/(1+\mathrm{e}^x)$, 每一句子的信息新颖度计算如下:

$$F_{IN}(s^T) = \zeta_2(KL(s^T \| HistoryTopicDis^T)) \tag{10}$$

根据每一句子的相关度和新颖度, 可得该句的总体评分为

$$Score(s^T) = \lambda \times F_{IR}(s^T) + (1-\lambda) \times F_{IN}(s^T) \tag{11}$$

为了避免每一时间段的摘要中出现冗余句子, 本文采用了 MMR 策略[23], 其中句子与句子之间的相似度计算方式如下:

$$\cos(s_1, s_2) = \frac{\sum\limits_{k=1}^{K} p(s_1|k) \times p(s_2|k)}{\sqrt{\sum\limits_{k=1}^{K} p^2(s_1|k)} \times \sqrt{\sum\limits_{k=1}^{K} p^2(s_2|k)}} \tag{12}$$

假设经过后验推理得到了 K 个主题 $\theta_1, \theta_2, \cdots, \theta_K$, 时间段 T 的主题分布 $TopicDis^T$ 计算如图 2 所示.

Initialize: $TopicDis^T = [\,]$
for each subtopic $\theta_k \in \{\theta_1, \theta_2, \cdots, \theta_K\}$,
　$TopicDis^T[k] = 0$
　for each s in s^T
　　$TopicDis^T[k] += p(k|s)$

图 2　时间段 T 的主题分布计算过程

Fig. 2　Calculation of topic distribution at time T

图 2 中, $p(k|s)$ 的计算如式 (6) 所示. 每一时间段之前的文档集的主题分布计算与图 2 类似.

2.3 摘要句子排序

要生成时序摘要, 需要将选择的句子按照一定方式排序. 本文从以下两个方面对句子进行排序得到摘要:

1) 根据选择句子来源文档的时间, 按照先后顺序排列;

2) 对于同一时间的句子, 根据其在文档中出现的相对位置排序, 位置靠前的排在前面. 具体为: 每一句子都对应一个取值区间为 $[0, 1]$ 的数值, 表示其在文档中的相对位置. 数值越小, 排序越靠前.

3 实验结果与分析

由于目前还没有标准的时序摘要评测数据集, 本文人工构建了评测数据集. 具体构建过程如下: 首先, 从互联网上随机选择若干突发性热门话题. 话题选择的标准是: 属于突发事件、规范性好、持续时间长. 选择突发事件是因为这类事件一般受关注度较高, 感兴趣用户较多; 规范性好则是考虑到互联网的海量数据中包含大量噪声信息, 很多数据本身很不规范, 处理起来难度较高; 持续时间长则是考虑到便于更好地评价时序摘要方法的性能. 然后, 邀请若干语言专家分别为几个话题人工生成摘要.

按照此构建过程, 本文的数据集实际包含了四个话题, 详细统计信息如表 2 和表 3 所示. 这四个话题均属于突发事件, 持续时间较长, 且分别来源于国内四大门户网站的新闻专题版块, 经过网站编辑的整理, 比较规范. 每个话题均邀请两名专家为其生成标准摘要.

表 2 数据来源信息

Table 2 Information of the data source

来源	篇数
新浪专题	380
腾讯专题	619
凤凰专题	620
网易专题	143

3.1 实验性能评价

本文采用文档摘要研究中的通用评价标准 Rouge[24] 对时序摘要的质量进行评价. 在 Rouge 评测指标中有多种子指标, 如 Rouge-N、Rouge-L、Rouge-W 等, 其中每一项评测指标都能产生出 3 个得分 (召回率, 准确率, F1 值). 下面以 Rouge-N 为例进行说明.

N 元语言模型的召回率 Rouge-N-R 为

$$\text{Rouge-N-R} = \frac{\sum\limits_{I \in GT} \sum\limits_{N\text{-}gram \in I} Count_{match}(N\text{-}gram)}{\sum\limits_{I \in GT} \sum\limits_{N\text{-}gram \in I} Count(N\text{-}gram)} \tag{13}$$

N 元语言模型的准确率 Rouge-N-P 为

$$\text{Rouge-N-P} = \frac{\sum\limits_{I \in CT} \sum\limits_{N\text{-}gram \in I} Count_{match}(N\text{-}gram)}{\sum\limits_{I \in CT} \sum\limits_{N\text{-}gram \in I} Count(N\text{-}gram)} \tag{14}$$

以上两者的 F1 值为

$$\text{Rouge-N-F} = \frac{2 \times \text{Rouge-N-P} \times \text{Rouge-N-R}}{\text{Rouge-N-P} + \text{Rouge-N-R}} \tag{15}$$

其中, N 是 N 元语言模型的长度, $N\text{-}gram \in GT$ 表示在标准答案摘要 GT 中出现的 N 元语言模型, $N\text{-}gram \in CT$ 表示在系统自动生成的摘要中出现的 N 元语言模型. $Count_{match}(N\text{-}gram)$ 是在候选文档摘要中和标准答案中都出现的 N 元语言模型数量, $Count\ (N\text{-}gram)$ 则表示仅出现在标准答案摘要或是系统自动生成的摘要中的 N 元语言模型数量. 受篇幅所限, 本文仅给出摘要的 Rouge-1-F、Rouge-2-F 和 Rouge-W-F 值. 由于标准的 Rouge 工具仅适用于英文评测, 本文对其进行了改进, 使其同样可以用于中文评测.

为方便评价生成摘要的质量, 本文参照多文档摘要评测任务的要求,限定每一话题摘要的最大长度

表 3 数据集的基本统计信息

Table 3 Detailed information of the data sets

话题	篇数	句子数	天数	标准摘要平均长度
1. 2013 年 12 月俄罗斯伏尔加格勒系列爆炸	137	1 628	7	650
2. 2013 年 7 月韩国波音客机旧金山坠毁	628	5 158	47	1 500
3. 2014 年 3 月昆明火车站恐怖袭击事件	270	2 661	12	850
4. 2013 年 4 月波士顿马拉松恐怖袭击	727	6 031	40	1 250

为标准摘要的平均长度, 然后评价所生成摘要的质量.

3.2　参数调优

在利用句子相关度和新颖度为句子评分时, 引入了参数 λ. 为了验证 λ 对摘要性能的影响, 本文以 0.05 为步长, 取 λ 在 [0, 1] 区间的变化值, 分别计算相应的 Rouge-1、Rouge-2 以及 Rouge-W 值. 图 3 展示了各度量值随 λ 增加的变化情况. 可以看到, 当 $\lambda = 0.35$ 时, 各度量值均达到最大, 并且从图 3 中可以看出:

1) 当 $\lambda = 0$, 即仅考虑句子的相关度时, 各度量值均比较低, 说明在时序摘要中跨时间段的冗余信息是存在的, 滤除冗余信息有助于提高摘要质量;

2) 当 $\lambda = 1$, 即仅考虑句子的新颖度时, 各度量值也出现了一定程度的下降, 这与实际情形相符, 即摘要首先要满足话题相关性.

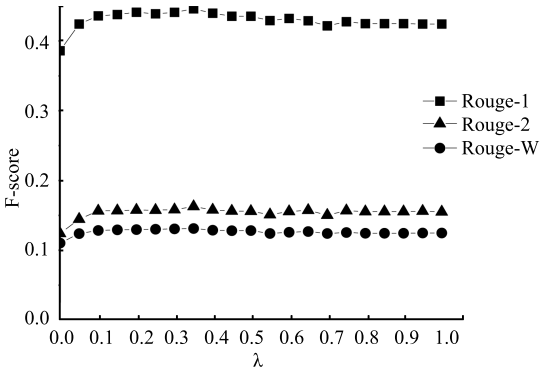


图 3　参数 λ 的影响

Fig. 3　Influence of parameter λ

3.3　与其他摘要生成方法的性能对比

本文选取文档摘要研究中的两种经典算法 LexRank[25] 和 Centroid[26] 作为 Baseline, 并将本文方法 (BDP) 与时序摘要研究中两种典型方法进行比较, 分别是基于宏微观重要性判别模型[14] 的方法 (Macro-Micro) 和基于 EHDP[10] 的方法 (EHDP). 总体性能对比如图 4 所示, 各方法在每一话题上的详细对比结果如表 4 所示. 由于 LexRank 算法和 Centroid 算法均未考虑时间信息, 本文对其抽取的句子按照句子出现的时间先后进行了排序.

LexRank 算法是文档摘要领域的经典算法, 首先根据句子之间的余弦相似度构建一个以句子为节点的连通图, 然后利用 PageRank 算法对节点进行排序, 最后选择排序靠前的句子作为摘要句.

Centroid 算法根据句子的 Centroid 评分、位置和与首句重复程度选择句子. Centroid 评分高、位置靠前且与首句重复信息少的句子会被选为摘要句.

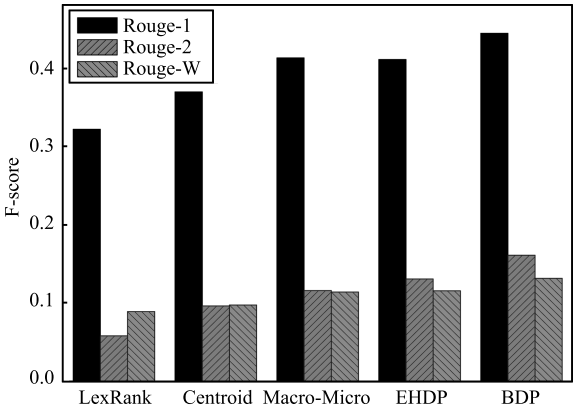


图 4　总体性能对比

Fig. 4　Overall performance comparison

Macro-Micro 算法首先选择重要时间段, 然后选择重要时间段内的句子作为该时间段的摘要.

EHDP 算法通过构建演化分层狄利克雷过程来利用句子内隐含的子话题信息. EHDP 方法没有考虑不同时间段之间的区别, Li 等[10] 在实验中假定每一时间段的摘要均不超过 50 个字. 如果机械按照 Li 等的实验设置, 会导致生成的摘要质量较差. 为更客观地衡量 EHDP 方法的性能, 本文在实现 EHDP 方法时, 根据每一话题的摘要总长度限制和时间段个数, 平均计算每一时间段摘要的长度.

从图 4 和表 4 可以看出, LexRank 和 Centroid 方法性能最差, 这是因为其仅根据词语特征判断句子重要性, 准确率低. 同时发现, Centroid 方法倾向于选择较长的句子, 因为长句子的 Centroid 评分较高, 更容易被选中为摘要句. 两者相比较, Centroid 方法性能较优, 这是因为其考虑的特征更全面.

Macro-Micro 方法对时间段进行了区分, 影响了召回率, 但是总体性能相较 LexRank 和 Centroid 方法提高不少. 该方法充分利用文档和句子在时间轴上的分布来判断时间段重要性并选择句子, 取得了较好性能, 这说明有必要对时间段进行区分. 但是该方法没有考虑跨时间段的冗余信息滤除, 生成的摘要中经常出现重复信息, 影响了该方法的性能. 另外, 其以句子所含事件触发词重要性的累加和作为句子重要性的判断依据, 并且倾向于选择包含新出现触发词的句子, 使得其容易受到噪声信息干扰. 例如, 话题 3 中的句子 “昨日下午, 继阿里木江 · 哈力克先生捐款后, 在广州摆地摊的维族男子卡哈尔 · 买买提也通过新疆都市报与本报联系, 希望能为 ‘3 · 01 严重暴力恐怖袭击事件’ 的伤者捐款, 经过了解后, 卡哈尔 · 买买提决定为伤情严重的石克香女士捐款 1 000 元”. 由于该句子包含多个新出现的动词使得其被 Macro-Micro 方法选为摘要句. 该句子虽然与话题相关, 但是属于次要信息, 标准摘要中并未包含

表 4 不同方法在四个话题上的结果

Table 4 Performance comparison on each topic

Topic	Method	Rouge-1	Rouge-2	Rouge-W
1. 俄罗斯伏尔加格勒爆炸	LexRank	0.31051	0.06526	0.09545
	Centroid	0.42807	0.14773	0.12828
	Macro-Micro	0.43800	0.14545	0.13228
	EHDP	0.44596	0.14178	0.12218
	BDP	0.47399	0.19010	0.14827
2. 韩国波音客机旧金山坠毁	LexRank	0.37693	0.06646	0.09337
	Centroid	0.36718	0.06350	0.08067
	Macro-Micro	0.43700	0.11519	0.11189
	EHDP	0.39111	0.12849	0.11393
	BDP	0.44160	0.12994	0.12443
3. 昆明火车站恐怖袭击	LexRank	0.26436	0.03571	0.06350
	Centroid	0.28905	0.04878	0.06066
	Macro-Micro	0.35191	0.07448	0.08501
	EHDP	0.37194	0.10576	0.09710
	BDP	0.39147	0.12082	0.10043
4. 波士顿马拉松恐怖袭击	LexRank	0.34955	0.05428	0.09428
	Centroid	0.31436	0.05420	0.07756
	Macro-Micro	0.38999	0.08717	0.10125
	EHDP	0.38666	0.13076	0.11918
	BDP	0.42917	0.16267	0.12900

类似信息.

EHDP 方法性能优于 Baseline, 说明挖掘隐含子话题信息有助于提高摘要质量. EHDP 方法在话题 2 上的 Rouge-1 值低于 Macro-Micro 方法, 原因是 Macro-Micro 方法仅生成重要节点的摘要, 而 EHDP 方法并未考虑节点的重要性. 由于话题 2 持续时间较长, 因此 EHDP 方法生成的摘要长度过长, 准确率相较 Macro-Micro 方法低, Rouge-1-F 值也受到了影响. 虽然 EHDP 方法在话题 2 上性能比 Macro-Micro 方法差, 但是其在其他 3 个话题上的性能几乎均优于 Macro-Micro 方法. 另外, EHDP 方法的 Rouge-2 和 Rouge-W 值优于 Macro-Micro 方法, 说明 EHDP 方法生成的摘要中包含的 2 元语言模型 (2-gram) 多数也在标准摘要中出现过. EHDP 方法和 BDP 方法均利用词语共现挖掘隐含子话题信息, 词语共现信息中包含着 2-gram, 因此挖掘隐含子话题有助于提高 Rouge-2 和 Rouge-W 值.

BDP 方法取得了最优性能. 虽然 BDP 方法与 EHDP 方法相比, 没有考虑摘要的连贯性; 与 Macro-Micro 方法相比, 没有考虑去除非重要时间段; 与 Centroid 方法相比, 没有在选择句子时考虑句子的位置等重要特征; 与 LexRank 方法相比, 没有考虑全局信息, 但是 BDP 方法生成的摘要最接近标准摘要, 表明 BDP 方法能更好地挖掘句子内的隐含语义信息, 并且根据句子的相关度和新颖度选择句子生成时序摘要是有效的. 另外, 经过观察生成的摘要, 有以下几点发现:

1) 考虑历史信息能够提高句子新颖度计算的准确率, 有效滤除跨时间段的冗余信息;

2) 由于是直接抽取句子生成摘要, 摘要中经常包含一些无用信息, 例如 “据 XX 报道”、“消息称” 等, 后续研究可以考虑进行句子压缩[27] 等优化;

3) 内容集中的话题摘要质量优于内容分散的话题摘要, 例如话题 1 和话题 4 的相关文档主要是报道相关事件本身的发展, 而话题 2 和话题 3 中有很多报道是关于记者对伤亡者家属和目击者的采访, 影响了摘要的质量;

4) 持续时间短的话题的摘要质量优于持续时间长的话题的摘要质量, 例如话题 1 的摘要质量优于其他持续时间较长的话题.

表 5 给出了利用本文方法生成的关于话题 1“俄罗斯伏尔加格勒爆炸” 的时序摘要.

4 总结

本文提出了一种基于 BDP 的时序摘要生成方法, 该方法直接利用文档集内的词语共现模式来挖掘句子的主题分布. 与传统主题模型相比, BDP 直接利用文档集内的词语共现模式发现主题,跳过了文

表 5 基于 BDP 生成的关于话题 “俄罗斯伏尔加格勒爆炸” 的时序摘要

Table 5 Temporal summary generated by BDP for “Explosions in the Russian city of Volgograd”

时序	话题摘要
2013-12-29	俄罗斯伏尔加格勒 1 号火车站于当地时间 29 日下午 12 : 45 时发生爆炸, 已致 18 死 40 多人受伤. 但报道并未提及爆炸发生的原因. 俄总统普京责成有关部门为受伤者提供一切必要的救助. 伏尔加格勒州政府将明年 1 月 1 日至 3 日定为哀悼日, 伏尔加格勒市政府宣布取消新年所有庆祝活动.
2013-12-30	俄罗斯国家反恐委员会在一份声明中说: “初步迹象显示, 爆炸由一名女性自杀式炸弹袭击者制造”. 据俄罗斯媒体 30 日报道, 俄国家反恐委员会表示, 伏尔加格勒州无轨电车恐怖袭击案的一种说法是安放在车厢内的炸弹被引爆.
2013-12-31	俄伏尔加格勒无轨电车遭袭. 策划 2011 年莫斯科多莫杰多沃机场自杀式袭击的伊斯兰武装势力头目乌马罗夫, 在今年 7 月公开叫嚣阻止索契冬奥会的举办. 伏尔加格勒发生恐怖爆炸事件后, 国际社会纷纷对恐怖袭击予以谴责, 并对遇难者家属以及俄罗斯政府和人民致以慰问.
2014-01-01	当地时间 2014 年 1 月 1 日凌晨, 俄罗斯总统普京乘坐专机抵达俄南部城市伏尔加格勒. 两天前, 这里曾经连续发生两起恐怖主义爆炸袭击. 这几天伏尔加格勒所有大规模的群众性新年庆祝活动都已经取消了.
2014-01-02	当地时间 2014 年 1 月 2 日, 俄罗斯伏尔加格勒为火车站爆炸袭击遇难者举行葬礼. 消息称: “警方逮捕了 700 余名行政违法人员、12 名被通缉的犯罪分子和 70 名嫌疑犯”. 消息指出, 在伏尔加格勒发生两起恐怖袭击后开展的旋风反恐行动中, 执法人员检查了全州 6 000 余处设施.
2014-01-03	俄内务部伏尔加格勒州内务总局 2 日称, 在新年第一天的 “旋风反恐” 行动中逮捕 700 余名违法者. 一名俄罗斯卫生部官员于当地时间 2 日表示, 又有 2 名在伏尔加格勒恐怖爆炸袭击中受伤的人员入院治疗, 这使得在地方诊所住院接受治疗的伤者人数升至 46 人.
2014-01-30	俄罗斯国家反恐委员会 30 日发布消息说, 在伏尔加格勒实施自杀式袭击的两名恐怖分子的身份已被查明. 此外, 俄罗斯强力部门工作人员 29 日在俄达吉斯坦共和国逮捕另外两人, 他们被指控涉嫌为恐怖分子实施袭击提供协助. 目前详细情况正在进一步调查.

档层, 避免了诸如句子等短文本的特征稀疏问题所造成的影响. 而且 BDP 能够自动发现文档集的主题数目, 易于对未知数据集进行建模. 在选择句子时, 本文通过计算句子的相关度和新颖度选择句子, 在计算句子新颖度时考虑了历史文档集, 减少了跨时间段的冗余信息. 虽然 BDP 模型应用于时序摘要取得了不错的效果, 但是 BDP 模型需要从整个文档集中抽取共现信息, 每当有新文档到来时需要重新对整个文档集进行建模. 未来的研究可以考虑在 BDP 中加入时序信息, 使得能够对摘要进行动态更新, 以适应在线处理的情形.

References

1 Yan R, Wan X J, Otterbacher J, Kong L, Li X M, Zhang Y. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In: Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China: ACM, 2011. 745−754

2 Yan R, Kong L, Huang C R, Wan X J, Li X M, Zhang Y. Timeline generation through evolutionary trans-temporal summarization. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK: ACL, 2011. 433−443

3 Tran G B, Tran T A, Tran N K. Leveraging learning to rank in an optimization framework for timeline summarization. In: Proceedings of the 36th Annual International ACM SIGIR Workshop on Time-aware Information Access. Dublin, Ireland: ACM, 2013. 433−443

4 Chieu H L, Lee Y K. Query based event extraction along a timeline. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK: ACM, 2004. 425−432

5 Xu S Z, Wang S S, Zhang Y. Summarizing complex events: a cross-modal solution of storylines extraction and reconstruction. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA: ACL, 2013. 1281−1291

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

7 Cao Jian-Ping, Wang Hui, Xia You-Qing, Qiao Feng-Cai, Zhang Xin. Bi-path evolution model for online topic model based on LDA. *Acta Automatica Sinica*, 2014, **40**(12): 2877−2886
(曹建平, 王晖, 夏友清, 乔凤才, 张鑫. 基于 LDA 的双通道在线主题演化模型. 自动化学报, 2014, **40**(12): 2877−2886)

8 Gao D H, Li W J, Zhang R X. Sequential summarization: a new application for timely updated twitter trending topics. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: ACL, 2013. 567−571

9 Huang L F, Huang L E. Optimized event storyline generation based on mixture-event-aspect model. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA: ACL, 2013. 726−735

10 Li J W, Li S J. Evolutionary hierarchical dirichlet process for timeline summarization. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: ACL, 2013. 556−560

11 Yan X H, Guo J F, Lan Y Y, Cheng X Q. A biterm topic model for short texts. In: Proceedings of the 22nd International World Wide Web Conference. Rio de Janeiro, Brazil: ACM, 2013. 1445−1455

12 Allan J, Gupta R, Khandelwal V. Temporal summaries of new topics. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, USA: ACM, 2001. 10−18

13 Lin F R, Liang C H. Storyline-based summarization for news topic retrospection. *Decision Support Systems*, 2008, **45**(3): 473−490

14 He Rui-Fang, Qin Bing, Liu Ting, Pan Yue-Qun, Li Sheng. Temporal multi-document summarization based on macro-micro importance discriminative model. *Journal of Computer Research and Development*, 2009, **46**(7): 1184−1191 (贺瑞芳, 秦兵, 刘挺, 潘越群, 李生. 基于宏微观重要性判别模型的时序多文档文摘. 计算机研究与发展, 2009, **46**(7): 1184−1191)

15 Chen C C, Chen M C. TSCAN: a content anatomy approach to temporal topic summarization. *IEEE Transactions on Knowledge and Data Engineering*, 2012, **24**(1): 170−183

16 Aslam J, Diaz F, Ekstrand-Abueg M, Pavlu V, Sakai T. TREC 2013 temporal summarization. In: Proceedings of the 22nd Text Retrieval Conference. Gaithersburg, USA: NIST, [Online], available: http://trec.nist.gov/pubs/trec22/trec-2013.html, January 1, 2015

17 Shou L D, Wang Z H, Chen K, Chen G. Sumblr: continuous summarization of evolving tweet streams. In: Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland: ACM, 2013. 533−542

18 Olariu A. Efficient online summarization of microblogging streams. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: ACL, 2013. 236−240

19 Olariu A. Hierarchical clustering in improving microblog stream summarization. In: Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics. Samos, Greece: Springer, 2013. 424−435

20 Zubiaga A, Spina D, Amigó E, Gonzalo J. Towards real-time summarization of scheduled events from twitter streams. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media. Milwaukee, USA: ACM, 2013. 319−320

21 Teh Y W, Jordan M I, Beal M J, Blei D M. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006, **101**(476): 1566−1581

22 Griffiths T L, Steyvers M. Finding scientific topics. *Proceedings of the National Academy of Science of the United States of America*, 2004, **101**(Suppl 1): 5228−5235

23 Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia: ACM, 1998. 335−336

24 Lin C Y, Hovy E. Automatic evaluation of summaries using N-gram co-occurrence statistics. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton, Canada: ACL, 2003. 71−78

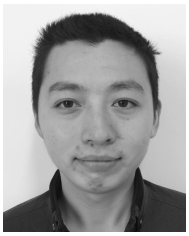
25 Erkan G, Radev D R. LexRank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 2004, **22**(1): 457−479

26 Radev D R, Jing H Y, Styś M, Tam D. Centroid-based summarization of multiple documents. *Information Processing and Management*, 2004, **40**(6): 919−938

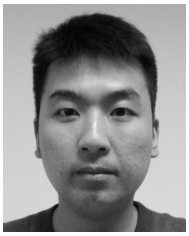
27 Li P, Wang Y L, Gao W, Jiang J. Generating aspect-oriented multi-document summarization with event-aspect model. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK: ACL, 2011. 1137−1146

席耀一 解放军信息工程大学信息系统工程学院博士研究生. 2011 年获得解放军信息工程大学硕士学位. 主要研究方向为自然语言处理. 本文通信作者.
E-mail: brian3333@163.com
(**XI Yao-Yi** Ph. D. candidate at the Institute of Information System Engineering, PLA Information Engineering University. He received his master degree from PLA Information Engineering University in 2011. His research interest covers natural language processing. Corresponding author of this paper.)

李弼程 解放军信息工程大学信息系统工程学院教授. 主要研究方向为文本分析与理解, 语音处理与识别, 图像/视频处理与识别, 信息融合.
E-mail: lbclm@gmail.com
(**LI Bi-Cheng** Professor at the Institute of Information System Engineering, PLA Information Engineering University. His research interest covers text analysis and understanding, speech/image/video processing and recognition, and information fusing.)



李天彩 解放军信息工程大学信息系统工程学院硕士研究生. 2012 年获得南京邮电大学学士学位. 主要研究方向为自然语言处理. E-mail: litc1125@126.com
(**LI Tian-Cai** Master student at the Institute of Information System Engineering, PLA Information Engineering University. He received his bachelor degree from Nanjing University of Posts and Telecommunications in 2012. His research interest covers natural language processing.)



黄山奇 65022 部队工程师. 2011 年获得解放军信息工程大学硕士学位. 主要研究方向为自然语言处理.
E-mail: luckyhsq@foxmail.com
(**HUANG Shan-Qi** Engineer at Unit 65022. He received his master degree from PLA Information Engineering University in 2011. His research interest covers natural language processing.)