文章编号：1003-0077(2018)09-0075-09

一种基于局部—全局主题关系的演化式摘要系统

吴仁守，刘 凯，王红玲

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：带有时间标志的演化式摘要是近年来提出的自然语言处理任务，其本质是多文档自动文摘，它的研究对象是互联网上连续报道的热点新闻文档。针对互联网新闻事件报道的动态演化、动态关联和信息重复等特点，该文提出了一种基于局部—全局主题关系的演化式摘要方法，该方法将新闻事件划分为多个不同的子主题，在考虑时间演化的基础上同时考虑子主题之间的主题演化，最后将新闻标题作为摘要输出。实验结果表明，该方法是有效的，并且在以新闻标题作为输入输出时，和当前主流的多文档摘要和演化摘要方法相比，在 Rouge 评价指标上有显著提高。

关键词：主题关系；PageRank；演化式摘要；多文档文摘

中图分类号：TP391　　**文献标识码**：A

An Evolutionary Summarization System Based on Local-global Topic Relationship

WU Renshou, LIU Kai, WANG Hongling

(Institute of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Evolutionary timeline summarization (ETS) for Internet News Event is a new task in natural language processing, which is a kind of multi-document summarization (MDS) in essence. According to the features of dynamic evolution, content relevance and information redundancy of Internet news event, this paper puts forward an evolutionary summarization method basing on local and global topic relations. First, the news event is divided into a number of different sub-topics. In the meantime, the basis of time evolution and the topic evolution between sub-topics are considered. Finally, headlines are extracted as summary. The experimental results show that this method is effective. Especially using news headlines as inputs or outputs brings significant improvements in the Rouge evaluation, compared with current popular method of multi-document summarization and evolution summarization.

Key words: topic relation; PageRank; evolutionary timeline summarization; multi-document summarization

0 引言

随着大数据时代的到来，互联网逐渐成为了人们获取和发布信息的主要渠道，互联网上关于热点新闻事件的报道与日剧增。当人们想要了解某一新闻事件时（例如，Egyptian Crisis），可以轻易在互联网上搜索到大量相关的报道，但是这些报道通常只是报道了这个新闻事件在某一时间段内的信息，且各个报道之间会有大量重复信息。面对海量的信息，人工逐一浏览归纳是非常耗时耗力的，为方便用户快速、全面地了解事件的发展，自动文摘成为一个有效手段。

传统多文档自动文摘把与事件相关的文档作为一个文档集合并为之生成摘要，集合中文档数目通常较少。面对互联网中大量相关且相似的文档，传统多文档自动文摘无法很好地工作。而且，由于传统多文档自动文摘没有考虑各个文档之间的时间和主题关系，很难让用户了解该事件的演化发展过程。与之相比，带有时间标志的演化式文摘（evolutionary timeline summarization，ETS）可以针对互联网上新闻事件的报道文档，按时间顺序抽取

收稿日期：2017-12-11　　**定稿日期**：2018-01-23

基金项目：国家自然科学基金（61402314）

出演化式摘要，为用户提供该事件全部发展过程，方便用户全面了解事件的前因后果和发展脉络，如表1所示。

随着时间推移，新闻话题的内容往往会发生变化，如何有效地组织这些大规模文档，生成主题在不同发展阶段的局部摘要，使其既能够提炼出主题的局部摘要信息，又能体现相邻时间段的主题演化，同时避免引入上一阶段的冗余信息，是演化式摘要面临的一个主要难题。由此，本文提出了一种基于局部—全局主题关系的演化式摘要方法，该方法将新闻标题作为文摘的候选句子，大大降低了数据量，并对事件进行主题分割，在考虑时间演化的基础上同时考虑子主题间的主题演化，最后通过一种改进的PageRank[1]算法将子主题和大主题相关联。该方法与以往方法的不同之处在于：以往方法通常通过抽取命名实体来追踪事件的演化，并且只考虑了时间维度或主题维度上的演化。本文除了考虑不同时间段的演化关系，还引入了子主题间的演化关系，并对传统的PageRank进行了拓展，利用句子、时间和主题三者相互强化来对句子打分排序。

表1 Wikipedia中关于Egyptian Crisis的部分带时间标签摘要

时间	摘要
5 January 2012	A prosecutor in the trial of Hosni Mubarak demanded that Mubarak be hanged, for the killing of protesters, during the 2011 uprising, that toppled his regime.
11 January 2012	The parliamentary elections were officially over.
24 January 2012	The leader of Egypt, Mohamed Hussein Tantawi, announced that the decades-old State of Emergency would be partially lifted, the following day.
1 February 2012	73 people were killed at a football game, in a stadium in Port Said.
17 March 2012	Pope Shenouda III died, at the age of 88. His passing greatly affected the entire nation of Egypt, and especially the Coptic Christian community.
24 March 2012	Numerous protesters took to the streets, angry that the football team El-Masry was banned for two more seasons, following the riots last month. The army then attacked the protesters. At least one person was killed, and at least 18 others were injured.

1 相关工作

传统多文档自动文摘(multi-document summarization, MDS)[2]是将同一主题下多个文本描述的主要信息按压缩比提炼出一个文本摘要的自然语言处理技术。根据文摘句选取方式的不同主要可分为两种：抽取型(extraction)文摘[3]和理解型(abstraction)文摘[4]。

作为多文档自动文摘的一种，演化式摘要为每个文档做上时间标记，然后按时间序列构成一个摘要，它的一个重要属性是动态演化性[5]。演化式摘要的动态演化性与话题检测与跟踪(TDT)任务中的话题演化研究类似，但又有所不同。TDT衡量的是同一个话题随时间推移表现出的动态性、发展性和差异性。演化式文摘通常针对单个新闻事件(或话题)，重点考虑内容演化，忽略强度演化。同时，不需要根据演化趋势做出预测，而需要根据演化趋势抽取代表句子生成摘要。

与时间有关的文摘技术最早由Allan[6]提出，通过抽取关键名词短语和命名实体来实现。Tran[7]也是通过抽取命名实体来追踪事件演化的，但是和上述方法不同的是，他利用了维基百科关于该事件的词条中实体的分布，并且综合考虑了实体在当前日期文档集合中的显著性(salience)和该实体在所有文档集中的信息性(informativeness)，据此抽取实体和它的上下文。Chieu[8]通过计算句子的新奇性(interest)和爆发性(burstiness)来抽取得分较高且日期不相连的句子作为摘要。不过这些方法都没有考虑新闻事件所特有的演化特性。Yan[9-10]使用基于图的方法，根据时间将句子映射到同一个平面，然后创建演化性文摘。其认为各个摘要组件既相互独立又相互联系，强调相关性、范围性、一致性和跨日期多样性，并通过构造一个最优化框架来平衡局部和全局的关系。其中Yan[10]通过将当前时间段文档集合附近的文档集，根据时间的间隔投影到当前文档集中来考虑文档集合之间的联系。该方法与本文提出的方法相似，但是它只考虑

了时间的演化而忽略了主题的演化。Li[11]把演化摘要任务看作主题演化的过程，利用层次狄利克雷过程为每一个日期文档集抽取出主题，捕捉主题演化的模式，通过考虑主题相关性、覆盖率和一致性等不同方面，抽取句子作为摘要。William[12]使用表示学习的方法把这个问题作为一个句子推荐任务。在纯文本语料库的基础上，利用了来自网络上排名最高的相关图像，使用卷积神经网络对图像进行建模，并提出了一种可扩展的低秩近似方法来学习新闻故事和图像的联合嵌入。Meena[13]使用与自然演化过程类似的遗传算法来优化线性搜索问题。

在中文方面，与 Tran[7]类似，宋俊等[14]提出面向实体的演化式多文档摘要生成方法，利用一个概率主题模型联合建模文档主题的演化和实体的参与情况，然后结合实体对句子进行评分和选择。徐伟等[15]利用词项强度和熵来确定代表性词项，然后基于内容覆盖性、时间分布性和传播影响力等三种指标构建出评价时间线摘要的综合评价指标，最后采用滑动窗口的方法，遍历时间轴上的微博消息序列，生成微博时间线摘要。

2　基于主题关系的演化式摘要方法

一个新闻事件通常包含了多个子主题[16]，每个子主题表现了这个新闻事件的某一个点。这些子主题往往是相互关联的，但是在一定程度上也是相互独立的，并不是所有子主题都和某个特定的子主题紧密关联。为了更好地刻画事件演化过程，我们分别从两种角度对事件主题进行建模：一种是局部的，基于子主题内部时间演化；另一种是全局的，基于子主题间主题演化。

如图 1 所示，我们将新闻标题集合 C 划分为 k 个子主题集$\{T_1, T_2, T_3, \cdots, T_k\}$。对于各个子主题 t_i，分别计算其对应的子主题内得分[Local(i)]与子主题间得分[Global(i)]，并生成子主题摘要。

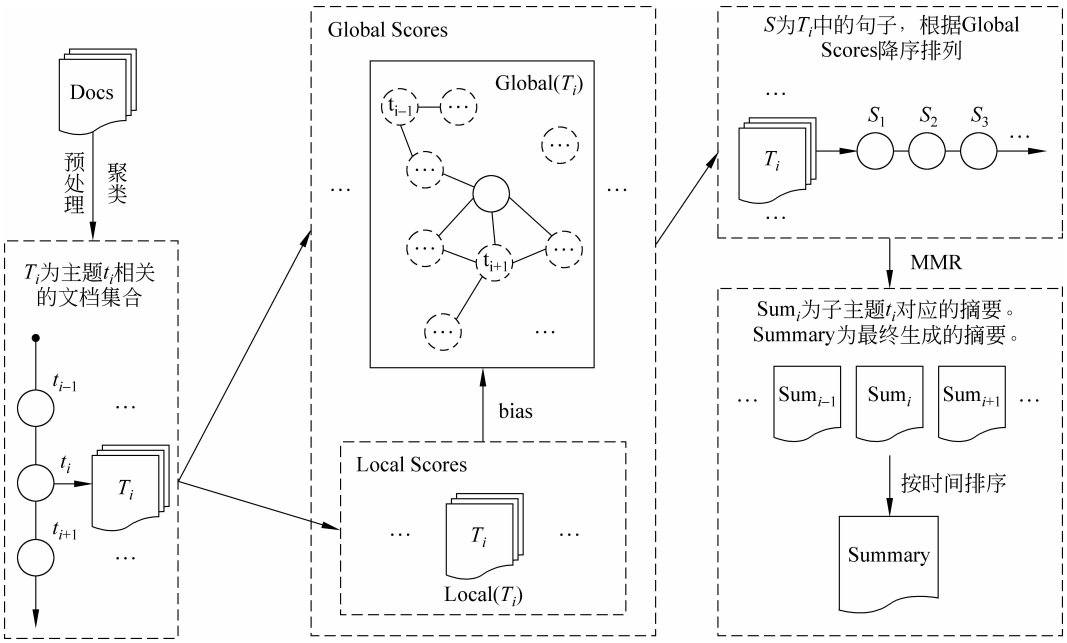

图 1　系统框架图

2.1　子主题内得分

在计算子主题内得分时，我们认为各个子主题是相互独立的。对于任意子主题 t_i，对应的标题集 T_i 中标题的主题基本相似，标题之间主题演化不明显，因此在计算子主题内得分时不考虑标题间的主题演化，而仅仅考虑其时间演化。

2.1.1　标题间时间距离

一般来说，如果两个标题之间的时间间隔越长，两个标题之间的联系就越弱，因此标题间的时间差异可以通过标题间时间距离来衡量。Rui Yan[10]利用高斯核函数将不同时间集中的句子映射到当前时间集来计算句子间的转移概率。与其类似但又有所不同，我们没有将不同时段的标题映射到同一时间

段上，而是通过圆核函数来计算两个标题之间的时间距离。圆核公式如式(1)所示。

$$\Gamma(ts_i - ts_j) = \begin{cases} \sqrt{1-\left(\frac{ts_i - ts_j}{\sigma}\right)2}, & \text{if} \mid ts_i - ts_j \mid \leqslant \sigma \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

其中 ts 表示时间戳，σ 为最大时间间隔。一般来说，σ 的最优设置可以根据新闻集而变化，因为句子可能在某些新闻主题中具有更广泛的语义范围，因此需要更高的 σ 值，反之亦然。

2.1.2 子主题内标题排序模型

通过对子主题 t_i 对应的标题集 T_i 构建有向图，可以使用普通的随机游走模型来计算子主题内各标题得分。设 $T_i = \{h_1, h_2, \cdots, h_n\}$，构建一个有向图 $G=(V,E)$，V 中的结点由 T_i 中的标题构成，结点 v_i 到 v_j 的边 e_{ij} 的权重由 v_i 到 v_j 的转移概率 p_{ij} 决定，如式(2)所示。

$$p_{ij} = \begin{cases} \frac{\Gamma(ts_i - ts_j) f_{ij}}{\sum_{\mid T_i \mid} \Gamma(ts_i - ts_k) f_{ik}}, & \text{if} \sum f \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

其中，f_{ij} 表示标题 h_i 和 h_j 对应的 TF-ISF 特征向量的余弦距离。

标题 h_j 得分通过模型中随机游走的访问概率来估计，该概率使用下列等式迭代计算，如式(3)所示。

$$S(h_j) = d\sum_i p_{ij} \times S(h_j) + (1-d) \tag{3}$$

其中转移概率 p_{ij} 在计算时已经归一化以满足马尔科夫属性，阻尼因子 $d=0.85$。当相邻两次迭代后，各个标题的得分差异小于 0.000 1 时，迭代停止。

2.2 子主题间得分

在计算子主题间得分时，我们认为各个子主题是相互关联的。因此，不仅要考虑各个标题之间的时间距离，还需要考虑各子主题间的主题差异。

2.2.1 子主题距离

子主题的主题特征向量由子主题中所有标题的特征向量求和取平均得到，并用余弦距离来衡量各个子主题间的差异。为了方便之后计算余弦相似性，在计算主题特征向量的同时，做了向量单位化，子主题 t_i 对应的主题特征向量 $\vec{t_i}$ 及子主题 t_i 和 t_j 之间的主题距离 d_{ij} 计算如式(4)、式(5)所示。

$$\vec{t_i} = \frac{\sum_{\mid T_i \mid} \vec{h_k}}{\sqrt{\left(\sum_{\mid T_i \mid} \vec{h_k}\right) \cdot \left(\sum_{\mid T_i \mid} \vec{h_k}\right)}} \tag{4}$$

$$d_{ij} = \vec{t_i} \cdot \vec{t_j} \tag{5}$$

2.2.1 子主题间标题排序模型

在为子主题 t_i 对应的标题集进行排序时，我们将其余子主题根据其与子主题 t_i 的距离映射到当前子主题中，并利用之前计算的子主题 t_i 内部排序结果为子主题 t_i 内的标题设置偏好。如图 1 所示，在计算 Global(i)时，实线圆代表当前计算的子主题 t_i，虚线圆代表映射到子主题 t_i 的其他子主题。为满足需求，我们对传统的 PageRank 算法进行了改变。

传统的 PageRank 算法表示如式(6)、式(7)所示。

$$\overrightarrow{\text{Rank}} = (1-\alpha)\boldsymbol{M} \times \overrightarrow{\text{Rank}} + \alpha\vec{p} \tag{6}$$

$$\vec{p} = \left[\frac{1}{N}\right]_{N\times 1} \tag{7}$$

其中 Rank 为 PageRank 值，$\boldsymbol{M}$ 为 $N \times N$ 的转移概率矩阵。Taher H[17]通过使用非均匀个性化向量 p 来增加某些类别的页面的影响，从而创建主题敏感的 PageRank。他认为偏置 p 涉及在计算的每次迭代中向适当的节点引入额外的等级，而不仅仅是在标准 PageRank 向量上执行的后处理步骤。与其类似，我们利用计算子主题内得分时得到的标题得分来修改 p，将子主题 t_i 内的局部特征融入到子主题间标题排序的全局建模中。

对于子主题 t_i，我们使用不均匀的 $p = v_{ji}$，如式(8)所示。

$$v_{ji} = \begin{cases} \frac{\beta}{N} + (1-\beta)S(h_i), & h_i \in T_i \\ \frac{\beta}{N}, & h_i \notin T_i \end{cases} \tag{8}$$

其中 β 为阻尼系数，设置为 0.8，$S(h_i)$ 为标题 h_i 在主题内的得分，N 为所有新闻标题的数目。

对于标题集 C 的所有标题，我们构建转移概率矩阵 M，p_{ij} 表示标题 i 到标题 j 的转移概率，如式(9)所示。

$$p_{ij} = \begin{cases} \frac{\Gamma(ts_i - ts_j) d_{ij} f_{ij}}{\sum_{\mid T_i \mid} \Gamma(ts_i - ts_k) d_{ik} f_{ik}}, & \text{if} \sum f \neq 0 \\ 0, & \text{if } T_i = T_j = T \end{cases} \tag{9}$$

有了转移矩阵 M 和偏置 p，就可以利用传统 PageRank 算法的求解过程进行求解。

2.3 摘要生成

根据各子主题对应的子主题间排序结果，分别从各个子主题中抽取一定数目的标题，并按照时间顺序输出作为摘要。各个子主题抽取的标题数目 η 由该子主题包含的标题数目 $|T_i|$ 以及总的标题数目 $|C|$ 来决定，具体计算如式(10)所示，最终生成摘要包含标题数目 n 如式(11)所示。

$$\eta = \frac{|T_i|}{\sqrt{|C|+b}} \tag{10}$$

$$n = \frac{\sum_{i=1}^{k}|C_i|}{\sqrt{|C|+b}} = \frac{|C|}{\sqrt{|C|+b}} \tag{11}$$

其中 k 为子主题数目，偏置 b 用于对最终生成摘要数目进行调整。当给定最终生成摘要包含标题数目 n 时，可以通过调整 b 的值进行控制。

一般的，子主题包含的标题数目越多，该子主题越重要，因此抽取的标题越多。当 η 小于 1 时，可以认为该子主题重要性较弱，不对其生成摘要。在冗余控制方面，利用最大边缘相关(MMR)[18]算法来去除冗余的句子。

3 实验与评价

3.1 数据集

我们利用 Giang Tran[19]论文中的数据集①，其中包含了埃及革命、叙利亚战争、也门危机和利比亚战争的四个长期事件。

数据集中的文章主要来源于 Google 搜索，针对用于创建参考摘要的新闻机构，构建了例如“埃及(革命或危机或起义或内战)”等问题，利用 Google 进行查询，并收集前 300 个答案。数据集中的参考摘要来源于包含 BBC、CNN 和 Reuters 等在内的多家知名通讯社出版的，由专业记者手动创建的对应事件的时间表。具体信息见表 2。

表 2　参考摘要概述

Story	# TL	# Timepoint	# GT-Date	# TL-Range	# a. sent	# News
Egypt Revolution	4	112	18	01/2011-07/2013	2	3 869
Libya War	7	118	51	02/2011-11/2011	2	3 994
Syria War	5	106	15	03/2011-08/2012	2	4 071
Yemen Crisis	5	81	22	01/2011-02/2012	2	3 600

注：参考摘要数量(# TL)，所有参考摘要时间点个数 # Timepoint，真实状况时间点个数(# GT-Date)，时间范围(# TL-Range)，每个参考摘要上每个日期的平均句子(# a. sent)，新闻文章数量(# News)

3.2 评价方法

ROUGE[20]是 Chin-Yew Lin 在 2004 年提出的一种自动摘要评价方法，被广泛应用于 NIST 组织的自动摘要评测任务中。ROUGE 基于摘要中 n 元词(n-gram)的共现信息来评价摘要，是一种面向 n 元词召回率的评价方法。基本思想为由多个专家分别生成人工摘要，构成标准摘要集，将系统生成的自动摘要与人工生成的标准摘要相对比，通过统计二者之间重叠的基本单元(n 元语法、词序列和词对)的数目，来评价摘要的质量。通过与标准人工摘要的对比，提高评价系统的稳定性和健壮性。该方法现已成为自动评价技术的通用标准之一。本文采用 ROUGE 中 ROUGE-1，ROUGE-2，ROUGE-L 和 ROUGE-SU4 的 F 值来对生成的摘要进行评价。

3.3 对照实验

LexPageRank[21]：是基于图排序的自动摘要算法，使用句子作为图结点，如果两个句子余弦相似度大于阈值则在这两个句子之间添加无向边，利用 PageRank 算法求解。其主要思想是：若一个句子与众多其他句子相似，那么此句话就可能是重要的。

Chieu：提出了“interest”和“burstiness”两种测量标准，认为在事件发生之后的一段时间内经常会在许多新闻文章中重复出现，并且有不同的更新和评论的事件是重要的。

ETTS：是迄今为止在新闻领域最好的无监督 TS 系统之一。它利用句子中的单词分布与整个语

① Available at http://www.l3s.de/*gtran/timeline/

料库中的单词分布以及相邻日期之间的相似性构造本地和全局摘要并进行优化组合。

3.4 实验设置

实验采用Java编程,运行服务器配置为3.40GHz Inter(R) Core(TM) i7-6700 CPU和16GB内存,使用Windows系统和JKD1.8.0_101环境。

我们将四个事件中的"也门危机"作为开发集,来对系统的各项参数进行调整。各个对比实验中的超参数均按照其对应论文中推荐的值进行设置,并根据本文数据集的大小及时间跨度等特点做了轻微的调整。在其他三个事件中进行交叉验证,将系统生成的自动摘要与人工生成的标准摘要利用ROUGE评价包(1.55版本)分别计算每次实验结果的ROUGE-1,ROUGE-2,ROUGE-L和ROUGE-SU4的F值,最后取平均值。

具体实验步骤如本文第二节。以下详细介绍文本预处理、子主题划分方法与参数设置。

3.4.1 文本预处理

我们对数据集中的特殊字符(例如@、#等)和长度小于4的标题进行过滤,并对单词进行词干提取以减少词表大小。最后,通过类似TF-IDF的TF-ISF(S为Sentences)技术将新闻标题转化为特征向量。

3.4.2 子主题划分

本文利用K-means聚类方法对新闻标题集合C进行子主题划分,子主题数目k值根据轮廓系数法[22]得到。通过枚举,令k从2到10取值,在每个k值上重复运行数次k_means,并计算当前k的平均轮廓系数,最后选取轮廓系数最大的值对应的k作为最终的集群数目。

3.4.3 参数设置

偏置b主要用于对生成摘要包含标题数目进行调整。为适应本文所用数据集,我们控制生成摘要包含标题数目在50个左右。因此,实验中b值统一设置为事件包含的总标题数目$|C|$的二分之一。

最大时间间隔σ主要用于对超出一定时间间隔的数据进行截断,减少干扰。为适应本文所用数据集,通过在"也门危机"事件数据集上对不同σ值进行实验,本文实验中σ值统一设置为一个月。

3.5 实验结果及分析

实验结果如表3所示(其中LGT为本文方法)。结果显示,本文提出的方法在三个事件中的各项测量指标均高于对比实验方法,说明本文提出的方法是有效的。和预想的相同,由于LexPageRank并没有考虑时间因素,所以其在三个事件中表现都为最差,而Chieu考虑了事件的演化性,所以效果优于LexPageRank。出乎意料的是:ETTS三个事件中的各项测量指标普遍高于Chieu,但是在叙利亚战争中ROUGE-2低于Chieu。通过比较两者生成的摘要,发现Chieu倾向于选择含有相似短语的标题作为摘要。当该短语是标准事件表中的重要短语时,Chieu获得了更高的ROUGE-2值。我们猜测这是由于Chieu算法本身造成的。它为句子集中具有高相似度的句子赋予高权重,而其自带的去冗余方法主要针对去除日期相近的句子。

表3 实验结果

系统	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
利比亚战争				
LexPageRank	0.271 51	0.059 90	0.268 81	0.087 09
Chieu	0.304 25	0.062 36	0.299 81	0.097 01
ETTS	0.340 17	0.068 65	0.334 17	0.116 48
LGT	**0.358 34**	**0.081 87**	**0.352 67**	**0.118 65**
叙利亚战争				
LexPageRank	0.287 00	0.059 16	0.279 57	0.088 50
Chieu	0.309 60	0.064 96	0.303 15	0.094 25
ETTS	0.319 61	0.063 51	0.314 38	0.098 09
LGT	**0.345 15**	**0.065 35**	**0.334 12**	**0.108 78**
埃及革命				
LexPageRank	0.269 90	0.072 31	0.265 53	0.092 22
Chieu	0.318 76	0.081 30	0.312 44	0.110 76
ETTS	0.351 21	0.089 26	0.344 41	0.120 33
LGT	**0.388 72**	**0.096 01**	**0.381 96**	**0.132 99**

针对利比亚战争,表4和表5分别列出了CNN人工编辑的部分摘要和本文方法抽取的部分摘要。结果显示,对于延续时间较长且在主题演化主线上的子主题,例如,"对利比亚实行禁飞区"和"北约对利比亚发动军事行动"等子主题,本文方法可以较好地识别并抽取出相关内容,说明本文的方法是有效的。但是,对于持续时间较短、偏离主题演化主线的事件,例如,"伊曼·奥贝迪(Eman al-Obeidy)事件"等,本文方法还是无法很好地识别出来。

表5显示,在冗余度控制方面,无论是在时间粒度上或是子主题粒度上,本文方法生成的摘要冗余度都很低。但是,本文方法生成的摘要会包含一些

评论性语句，例如，“What can be done to end the crisis in Libya?”等，这些语句通常并没有涉及具体的事件，不适合作为时间标签摘要。我们猜想，抽取到这些句子的原因可能是这些评论性句子中通常包含多个该事件下的主题词，例如，“libya”“crisis”等，导致这些句子获得了高得分。

表 4　CNN 针对利比亚战争人工编辑的时间标签摘要（节选，2011 年 3 月）

时间	摘要
2011-03-01	The United Nations General Assembly adopts a resolution to oust Libya from its seat on the 47-member Human Rights Council. 联合国大会决议将利比亚从 47 人的人权理事会席位上赶下台。
2011-03-07	NATO begins round-the-clock surveillance flights of Libya as it considers various options for dealing with escalating violence there. 考虑到各种应对暴力升级的情况，北约开始对利比亚进行二十四小时的飞行监视。
2011-03-17	The United Nations Security Council votes to impose a no-fly zone over Libya and take “all necessary measures” to protect civilians. 联合国安理会投票决定对利比亚实行禁飞区，并采取“一切必要措施”保护平民。
2011-03-18	Libyan Foreign Minister Moussa Koussa says the country has decided on “an immediate cease-fire and the stoppage of all military operations.” But sources inside Libya say violence continues. 利比亚外长穆萨库萨说，该国已决定“立即停火，停止一切军事行动。”但是利比亚内部消息人士说暴力仍在继续。
2011-03-19	French，British and American military forces begin the first phase of operation “Odyssey Dawn”，aimed at enforcing the no-fly zone. 法国、英国和美国军队开始了“奥德赛黎明”行动的第一阶段，旨在执行禁飞区。 More than 110 Tomahawk missiles fired from American and British ships and submarines hit about 20 Libyan air and missile defense targets，U. S. Vice Adm. William Gortney says at a Pentagon briefing. 美国副总理威廉·高特尼（William Gortney）在五角大楼简报会上说，美国和英国的舰艇和潜艇发射的 110 枚“战斧”导弹击中了大约 20 个利比亚的空中和导弹防御目标。
2011-03-20	Gadhafi，speaking on Libyan state TV，says the U. N. charter provides for Libya’s right to defend itself in a “war zone.” 卡扎菲在利比亚国家电视台发表谈话时表示，联合国宪章规定利比亚有权在“战区”进行自卫。
2011-03-24	NATO agrees to take command of the mission enforcing a no-fly zone over Libya. 北约同意执行对利比亚实施禁飞区的任务。
2011-03-26	A Libyan woman（*Eman al-Obeidy*）with bruises all over her body bursts into a Tripoli hotel housing international journalists，shouting that she was taken from a checkpoint and held for two days while 15 of Gadhafi’s militiamen beat and raped her. 一名全身瘀伤的利比亚女子（*Eman al-Obeidy*）闯入了一家收容国际记者的黎波里旅馆，并叫喊着她从检查站被带走了两天，而此期间卡扎菲的 15 名民兵殴打并强奸了她。

表 5　LGT 生成关于 Egyptian Crisis 的部分带时间标签摘要（节选，2011 年 3 月）

时间	摘要
2011-03-01	Libya Will Obama Order U. S. Military Intervention? 奥巴马下令美国将军事干预利比亚？
2011-03-04	What can be done to end the crisis in Libya? 可以做些什么来结束利比亚的危机？
2011-03-10	Libya’s war Rebels flee Gaddafi’s force as NATO vetoes nofly zone without UN resolution. 因为北约在没有联合国决议的情况下否决了禁飞区，所以利比亚战争的反动者逃离了卡扎菲武装力量。
2011-03-17	Libya crisis Britain，France and US prepare for air strikes against Gaddafi. 面对利比亚危机，英、法、美准备对卡扎菲进行空袭。

续表

时间	摘要
2011-03-18	Libya crisis World strikes back on Gaddafi as UN votes to protect Libyan rebels. 面对利比亚危机,联合国表决决定保护利比亚反叛分子,世界反击卡扎菲。
2011-03-19	Military action launched against Libyan forces. 发起军事行动对抗利比亚武装力量。
2011-03-24	Coalition agrees to put NATO in charge of no-fly zone in Libya. 联盟同意让北约负责利比亚禁飞区。

4 结束语

演化式摘要作为多文档自动文摘的一种,它在传统多文档摘要的基础上需要额外考虑事件随时间变化的演化特性。为此,本文提出了一种基于主题和时间变化的演化式摘要方法,其分别考虑了子主题内部和子主题间的主题和时间关系,并通过变种的 PageRank 算法将两者联系起来。实验结果表明,该方法与现有方法相比在 ROUGE 值上有较大提升。未来,我们将通过句法结构分析、单词词性等文本特征来判断句子是否属于评论性句子,避免这些不涉及具体事件的句子在摘要中出现。另外,我们还将尝试不同的文本表示方法,例如,LDA、Word Embedding 等,并考虑将时间和主题等特征加入文本表示的向量中。

参考文献

[1] Page L. The PageRank citation ranking: Bringing order to the web[J]. Stanford Digital Libraries Working Paper,1998,9(1): 1-14.

[2] 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.

[3] Wong K F,Wu M,Li W. Extractive summarization using supervised and semi-supervised learning[C]//Proceedings of the COLING 2008,International Conference on Computational Linguistics,Proceedings of the Conference,UK,2008: 18-22.

[4] Nallapati R,Zhou B,Santos C N D,et al. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond[C]//Proceedings of the Signll Conference on Computational Natural Language Learning,2016: 280-290.

[5] Jo Y,Hopcroft J E,Lagoze C. The web of topics: discovering the topology of topic evolution in a corpus [C]//Proceedings of the International Conference on World Wide Web, WWW 2011, Hyderabad, India, DBLP,2011: 257-266.

[6] Allan J,Gupta R,Khandelwal V. Temporal summaries of news topics[C]//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA. DBLP,2001: 10-18.

[7] Tran T A,Niederee C,Kanhabua N,et al. Balancing novelty and salience: Adaptive learning to rank entities for timeline Summarization of high-impact events[C]//Proceedings of the ACM International on Conference on Information and Knowledge Management. ACM,2015: 1201-1210.

[8] Hai L C,Lee Y K. Query based event extraction along a timeline[C]//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval,2004: 425-432.

[9] Yan R,Wan X,Otterbacher J,et al. Evolutionary timeline summarization: A balanced optimization framework via iterative substitution[C]//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval,2011: 745-754.

[10] Yan R,Kong L,Huang C,et al. Timeline generation through evolutionary trans-temporal summarization [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2011.

[11] Li J,Li S. Evolutionary hierarchical dirichlet process for timeline summarization[C]//Proceedings of Meeting of the Association for Computational Linguistics,2013: 556-560.

[12] Wang W Y,Mehdad Y,Radev D R,et al. A Low-Rank approximation approach to learning joint embeddings of news stories and images for timeline summarization[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2016: 58-68.

[13] Meena Y K,Gopalani D. Evolutionary algorithms for extractive automatic text summarization [J]. Procedia Computer Science,2015(48): 244-249.

[14] 宋俊,韩啸宇,黄宇,等. 一种面向实体的演化式多文

档摘要生成方法[J]. 广西师范大学学报(自然科学版),2015,33(02): 36-41.

[15] 徐伟,赵斌,吉根林. 基于滑动窗口的微博时间线摘要算法[J]. 数据采集与处理,2017,32(3): 523-532.

[16] Wang H,Zhou G. Topic-driven multi-document summarization[C]//Proceedings of International Conference on Asian Language Processing. IEEE, 2011: 195-198.

[17] Haveliwala T H. Topic-sensitive pageRank: A context-sensitive ranking algorithm for web search[J]. Knowledge and Data Engineering IEEE Transactions on,2003,15(4): 784-796.

[18] Carbonell J,Goldstein J. The use of MMR,diversity-based reranking for reordering documents and producing summaries [C]//Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,1998: 335-336.

[19] Tran G,Alrifai M,Herder E. Timeline summarization from relevant headlines[M]. Advances in Information retrieval. Springer International Publishing, 2015: 245-256.

[20] Lin C Y,Hovy E. Automatic evaluation of summaries using N-gram co-occurrence statistics[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics,2003: 71-78.

[21] Erkan G, Radev D R. LexPageRank: Prestige in multi-document text summarization[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A Meeting of Sigdat,A Special Interest Group of the ACL,Held in Conjunction with ACL 2004,25-26 July 2004,Barcelona,Spain. DBLP,2004: 365-371.

[22] 夏士雄,李文超,周勇,等. 一种改进的 K-means 聚类算法[J]. 东南大学学报(英文版),2007,23(3): 435-438.

吴仁守(1995—),硕士研究生,主要研究领域为自然语言处理、信息检索。
E-mail: 20165227037@stu.suda.edu.cn

刘凯(1992—),硕士,主要研究领域为自然语言处理、信息检索。
E-mail: 20154227024@stu.suda.edu.cn

王红玲(1975—),博士,副教授,主要研究领域为自然语言处理、信息检索。
E-mail: hlwang@suda.edu.cn