文章编号：1003-0077(2019)04-0093-08

基于查询的新闻多文档自动摘要技术研究

王凯祥，任 明

（中国人民大学 信息资源管理学院，北京 100872）

摘 要：针对新闻文本领域，该文提出一种基于查询的自动文本摘要技术，更加有针对性地满足用户信息需求。根据句子的 TF-IDF、与查询句的相似度等要素，计算句子权重，并根据句子指示的时间给定不同的时序权重系数，使得最近发生的新闻内容具有更高的权重，最后使用最大边界相关的方法选择摘要句。通过与基于 TF-IDF、TextRank、LDA 等六种方法的对比，该摘要方法 ROUGE 评测指标上优于其他方法。从结合评测结果及摘要示例可以看出，该文提出的方法可以有效地从新闻文档集中摘取核心信息，满足用户查询内容的信息需求。

关键词：自动文本摘要；基于查询的摘要；新闻文本；分布式表示

中图分类号：TP391 **文献标识码**：A

Query-based Multi-document Automatic Summarization of News

WANG Kaixiang, REN Ming

(School of Information Resource Management, Renmin University of China, Beijing 100872, China)

Abstract: This paper proposes a query based automatic text summarization method, which is targeted to meet users' information needs of news. It assigns the weight of the sentence according to the TF-IDF, the similarity of sentence to the query, and the time of the sentence indicating (with a bias favoring the recent news). Finally, the method of the Maximal Marginal Relevance is used to select the summary sentence. Compared with six existing methods, the method proposed in this paper is superior in terms of ROUGE.

Keywords: automatic text summarization; query-based summary; news text; distributed representation

0 引言

在人工智能技术日新月异、互联网技术飞速发展、人们信息需求不断提升的今天，信息传播渠道丰富多样，人们每天都会接收大量的信息，从这些海量信息中找出自己所需信息，需要花费大量的时间和精力。自动文本摘要技术的出现，可以帮助人们节省大量阅读时间，在相同的时间内获取更多的有效信息。基于查询的自动文本摘要技术可以对用户感兴趣的、主动查询的内容进行摘要，更加有针对性地满足用户的信息需求，方便用户更快更准确地获取到所需的内容，提高阅读效率，提升阅读体验。

自动文本摘要方法主要有两大类：生成式(abstractive)和摘取式(extractive)，生成式需要在语义理解的基础上，在词语级别上生成摘要。摘取式是通过分析文本统计特征、潜意语义特征等，在句子或段落级别上生成摘要。其中摘取式摘要方法从方法技术上分主要包括基于统计信息、基于机器学习、基于主题模型、基于图模型等方法。

基于查询的自动摘要技术主要在通用自动摘要算法基础上，针对面向查询的特点，对相关技术进行了适用性改进。在基于图模型的自动摘要方法上，使用流排序算法可以计算加入查询节点后，权重在图中传播后的各个节点的权重。Cai 和 Li 在流排序的基础上，增加了主题层的排序[1]。Canhasi[2] 基于 PageRank 构建了在句子、查询句、段落、文档、框架五个层面的图模型，进一步提高了模型效果。超图模型可以使传统的图模型结构连接超过两个句子，

收稿日期：2018-05-29 **定稿日期**：2018-08-17

基金项目：国家自然科学基金(71402186)

降低复杂度。Xiong 和 Ji[3]结合主题模型获得主题分布,使用超图模型获得词与主题、句子与句子的主题分布,通过节点增强和随机游走模型对句子进行排序。Zheng 等在此基础上增加了概念层[4]。在基于聚类的自动摘要方法上,在根据句子或词语之间的相似度对句子进行聚类时,会加入语义信息[5]和多种特征[6-7],提高相似度计算的准确率和聚类效果,其中聚类方法的改进也会提升摘要效果。Naveen 和 Nedungadi 使用 PHA-Clustering Gain 与 K-Means 结合方法改进了聚类方法[8]。Yang[9]基于 HLDA 并结合 n-gram 模型,提出了一种考虑上下文关系的主题模型。聚类方法与图模型的结合在多文档摘要中表现较好。Sun 等[10]在聚类之后构建两层图模型,通过寻找最优路径的方式提取摘要。

在基于机器学习的自动文本摘要方法上,通常通过提取与查询语句相关的特征[11],如句子位置、长度、与查询句子的相似度、TF-IDF 等特征[12-13],以优化摘要结果。随着标注数据的增多和深度学习的发展,神经网络模型在生成式摘要的应用上逐渐增多[14-16],但其在语义可读性上的表现有待提高。

词语的向量表示是通过相关模型将每个词语转换成唯一的特征向量,Mikolov 等针对词语的分布式表示[17],提出了通过神经网络语言模型获得其分布式向量表示的 Word2Vec 方法,可以通过词之间的距离来判断它们之间的语义相似度,该方法在词向量降维、语义分析、相似度计算等方面均有较好的表现。本文使用 Word2Vec 的方法计算词语之间的相似度,进而得出句子相似度。

句子的选择通常需要满足三个条件:①所选句子对文摘信息量的增加贡献度尽量高;②使文摘的信息冗余度尽量低;③所选句子数量满足摘要对句子或词语数量的限制要求。最大边界相关法(maximal marginal relevance,MMR)[18]可以从候选摘要句子集中选择出句子权重高同时使摘要集冗余度低的最优句子,满足摘要句选择的要求[19]。

1 基于查询的自动文本摘要

本文基于查询的新闻多文档自动摘要方法的设计,从主要流程上包括数据集的获取、文档预处理、句子权重及时序权重系数计算、句子相似度计算、句子选择几个部分,具体流程如图 1 所示。

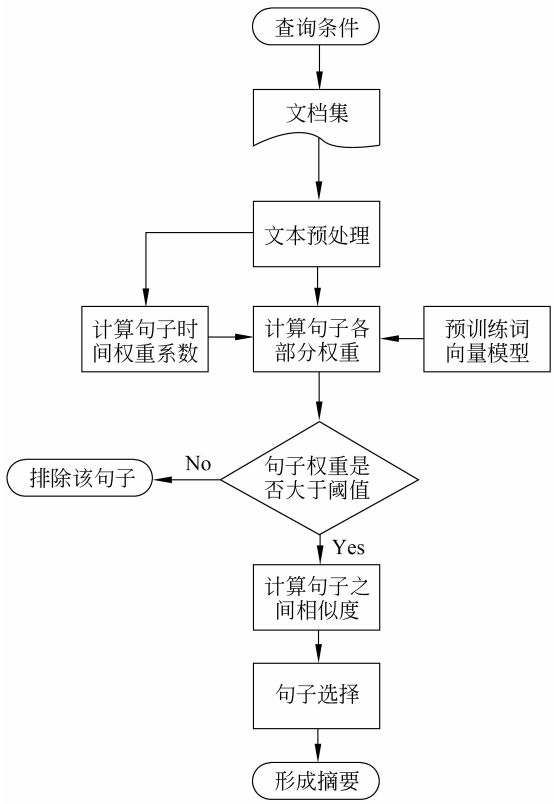


图 1 基于查询的新闻多文档自动摘要设计流程

1.1 句子相似度计算

本文使用 gensim 工具包中的 Word2Vec 训练获取的词向量计算句子相似度,由于对于某个检索条件下的新闻文档集,其各个文档及文档中句子的相似度很高,使用词向量模型可以更加准确地区分各个词语之间的语义差别。词向量模型的训练结果得到的是每个词语的向量值,将这些向量映射到维度空间中,就得到了词向量的空间模型,词语之间的相似度值可以使用两个词语在空间模型中的余弦相似度表示[20],如式(1)所示。

$$\cos(\boldsymbol{w}_a,\boldsymbol{w}_b)=\frac{\sum_{k=1}^{n}(w_{a_k}\times w_{b_k})}{\sqrt{\sum_{k=1}^{n}(w_{a_k})^2}\times\sqrt{\sum_{k=1}^{n}(w_{b_k})^2}} \tag{1}$$

其中:$\boldsymbol{w}_a$、$\boldsymbol{w}_b$ 表示词向量,n 表示词向量的维数,w_{a_k} 表示 $\boldsymbol{w}_a$ 向量的第 k 维的值。w_{b_k} 表示 $\boldsymbol{w}_b$ 向量的第 k 维的值。

句子是由一个个词语组成的,所以句子之间的相似度可以在词语相似度的基础上计算得到。我们知道两个句子中相似的词语越多其相似度应该越高,当两个句子完全一样时其相似度为 1。同时为了避免长句子的相似度过高,减弱长句子在词语数

量上的优势，这里采用先求和再求平均数的方法，计算词语相似度的平均值。因此两句子相似度计算，如式(2)所示。

$$\mathrm{sim}(s_i,s_j)=\frac{\sum_{w_i\in s_i}\max_{w_j\in s_j}(\cos(\boldsymbol{w}_i,\boldsymbol{w}_j))+\sum_{w_j\in s_j}\max_{w_i\in s_i}(\cos(\boldsymbol{w}_i,\boldsymbol{w}_j))}{L_{s_i}+L_{s_j}} \tag{2}$$

其中，$\mathrm{sim}(s_i,s_j)$ 表示句子 s_i 与句子 s_j 的相似度，w_i 表示 s_i 中的词语，w_j 表示 s_j 中的词语，$\cos(w_i,w_j)$ 为 w_i,w_j 两个词语的向量空间余弦相似度，L_{s_i},L_{s_j} 为 s_i,s_j 中包含词语的数量。

1.2　句子权重的计算

本文在句子权重的计算中主要考虑以下五部分的因素。

(1) TF-IDF 得分。词频(term frequency，TF)指的是某一个给定的词语在该文件中出现的频率。逆向文件频率(inverse document frequency，IDF)是一个词语普遍重要性的度量。其计算如式(3)所示。

$$\mathrm{TF}_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}},\quad \mathrm{IDF}_i=\log\frac{N_D}{N_{\{j:w_i\in d_j\}}} \tag{3}$$

其中，$n_{i,j}$ 是该词 w_i 在文件 d_j 中的出现次数，N_D 表示语料库中的文件总数，$N_{\{j:w_i\in d_j\}}$ 表示包含词语 w_i 的文件数目。在计算句子的 TF-IDF 得分时，为了避免长句子的得分偏高，使用句子所包含词语的 TF-IDF 平均值来表示句子的 TF-IDF 值，如式(4)所示。

$$\mathrm{weight}_{\mathrm{tfidf}}=\frac{\sum_i \mathrm{TF}_{i,j}\times\mathrm{IDF}_i}{L_s} \tag{4}$$

其中，$\mathrm{weight}_{\mathrm{tfidf}}$ 表示句子的 TF-IDF 得分，L_s 表示句子长度。

(2) 位置权重。由于新闻文本结构的倒金字塔特性，整篇新闻的最核心内容往往会放在首段或者首句进行说明。通过对大量新闻文章的调研发现，新闻文章为了吸引读者兴趣，使用首段首句引出后面所要表达的核心内容，在首段中首句之后的句子仍然表达的是总结性的内容，所以此权重计算如式(5)所示。

$$\mathrm{weight}_{\mathrm{pos}}=\begin{cases}1,\text{if 位于首段}\\0,\text{else 非首段}\end{cases} \tag{5}$$

(3) 与标题的相似度。新闻文章的标题通常会以最凝练的语言概括整篇文章的主要内容，所以句子与标题的相似度，可以体现出该句子与文章主要内容的相关程度，如式(6)所示。

$$\mathrm{weight}_{\mathrm{t}}=\mathrm{sim}(s,T) \tag{6}$$

其中，s 为句子，T 为标题。

(4) 与查询的相似度。查询语句体现了用户所要了解的信息范围，句子与查询语句的相似度越高表示该句子更有可能是用户想要阅读的内容，如式(7)所示。

$$\mathrm{weight}_{\mathrm{q}}=\mathrm{sim}(s,Q) \tag{7}$$

其中，Q 代表查询。

(5) 线索词权重。线索词是指“总而言之”“总的来说”等概括性的指示词语，包含线索词的句子通常是对其他文章内容的总结，会包含更多的信息，在权重设置上应给予更高权重：

$$\mathrm{weight}_{\mathrm{x}}=\begin{cases}1,\text{if 包含线索词}\\0,\text{else 不包含线索词}\end{cases}$$

句子权重由以上五个部分组成，为了平衡各部分权重得分的分布，为每部分权重引入了权重系数，该权重系数由两部分组成：归一化系数和经验权重。如式(8)所示。

$$\lambda=\alpha\times\varepsilon \tag{8}$$

其中归一化系数 ε 是通过计算已知文档集上五种权重的分布，对其进行归一化后得到的系数，经验系数 α 是根据实验分析，调优后的参数。

句子最终权重值为权重系数与各部分权重值得乘积之和，即：

$$W_{\mathrm{group}}(s)=\lambda_{\mathrm{tfidf}}\,\mathrm{weight}_{\mathrm{tfidf}}+\lambda_{\mathrm{pos}}\,\mathrm{weight}_{\mathrm{pos}}+\lambda_{\mathrm{t}}\,\mathrm{weight}_{t}+\lambda_{\mathrm{q}}\,\mathrm{weight}_{\mathrm{q}}+\lambda_{\mathrm{x}}\,\mathrm{weight}_{\mathrm{x}} \tag{9}$$

其中，λ 为各部分的权重系数，W_{group} 为句子 s 五个权重要素结合后的组合权重值。

1.3　时序权重系数

新闻报道的一大特点是讲求时效性，同时较新的文章会包含以前的新闻事件的介绍；用户在搜索某个新闻内容时通常也是为了获得最新的新闻进展。所以较新的新闻内容更符合用户的信息需求，在计算句子权重时应考虑新闻时效的影响。

由于摘要结果是在句子维度上对句子进行的选择，所以对时间区分的维度应该也在句子维度上，即要确定每个句子所对应的时间。通过对大量新闻文章的句子时间的分析发现：新闻文章的段落较短，同一段落的句子往往只表达一层含义或一个观点，在未出现时间标识词时，往往表示同一时间。本文

提取句子时序特征的流程如图2所示。

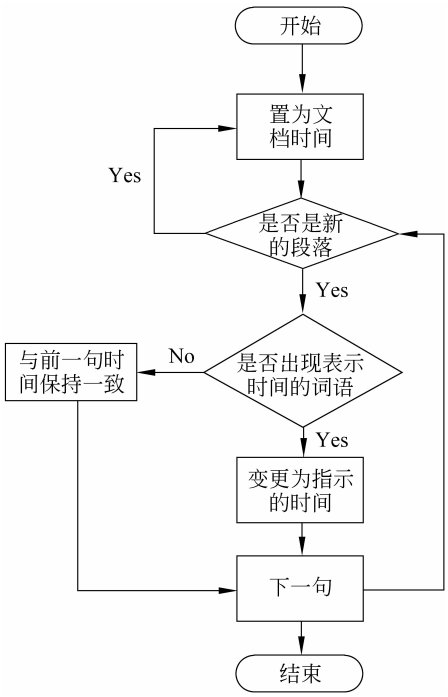


图2 句子时序特征提取流程

由于新闻文章的时效性，在提取摘要时，距离现在越远的句子，其权重应该越低。同时为了避免因句子间隔时间不等导致的权重波动问题，这里采用的是相对时间，也就是句子在时间排序上的位次，位次越靠前权重值越高。λ_{time} 为时序衰减系数。

由此，对文当中任意一个句子 s 的权重值如式(10)所示。

$$W(s) = \lambda_{\text{time}} W_{\text{group}}(s) \tag{10}$$

在时序衰减函数的选择上，需使衰减系数的取值范围为0～1，本文分别对比了三种衰减函数方式的效果：(1)常数型：α，即衰减系数是不随时间发生变化的常数。(2)线性型：$1-\alpha\frac{n}{N}$，即衰减系数是随时间线性变化的一次函数形式。(3)指数型：$e^{-\alpha\frac{n}{N}}$，即衰减系数是随时间变化的指数形式。其中 α 为经验系数，N 为文档集总句子数，n 为该句子在该文档集中按时间倒序排序后的排序值。

对比三种形式的衰减函数，在标注数据上，选择不同文摘比例上的 F 值如图3所示，其中 α 取默认值1。

从图3中可以看出，三条曲线随着文摘比例的增大而增大，这是由于人工摘要的句子数是一定的，当摘要比例增大时，机器摘要的正确句子数则会逐渐增加，准确率和召回率也随之增加。通过对比可以发现指数型的衰减函数，在时间序列的处理上有更好的效果。

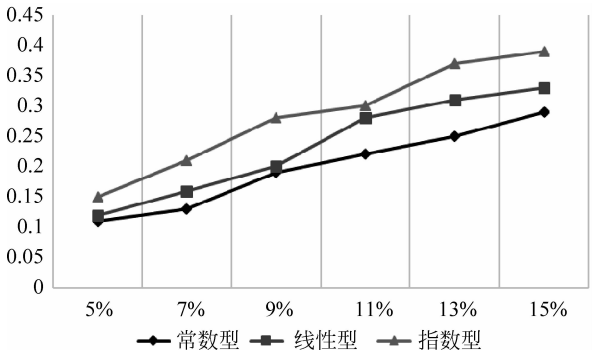


图3 不同衰减函数效果对比

在15%文摘比例下，对于常数型、线性型、指数型衰减函数，选择不同经验系数 α 取值的 F 值对比，如图4所示。

从图4中可以看出：(1)对于常数型衰减函数，其参数大小对句子权重的相对大小没有影响，所以 F 值不随其变化；(2)对于线性型、指数型衰减函数：当系数 α 逐渐增大时，F 值逐渐变高，这是由于当 α 过小时，衰减函数的取值与常数型接近，不能体现出时序衰减的特性；当 α 继续增大时，则会使时间较新句子权重偏大，使摘要集里时间较新的句子增多，导致 F 值降低；(3)经过调优后可以看出指数型的最大值要大于线性型的最大值，所以指数型衰减函数要优于线性衰减函数。

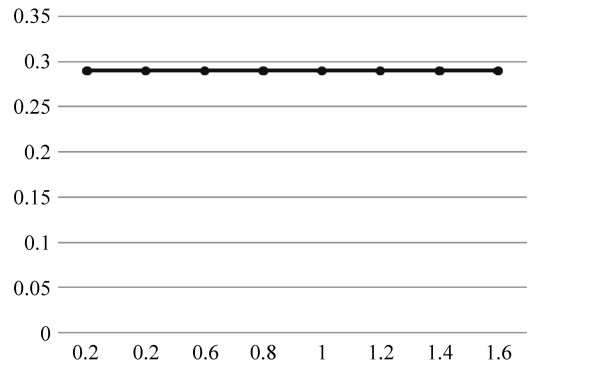
2 实验过程与结果分析

2.1 数据准备

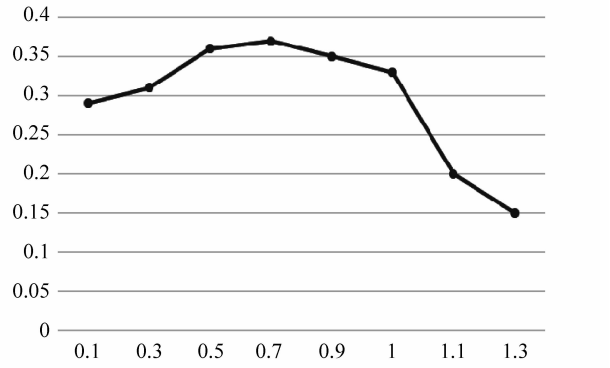
研究的文本对象是中文新闻文本，研究的主题是基于查询的自动文本摘要，当前在中文领域没有适合本研究主题的标注语料集，同时为了结合使用的实际情况，采用基于新闻网站搜索引擎结合语句查询的方法，通过爬虫抓取查询结果文档，组成文档集。实验选取的新闻网站为光明网，一是由于光明网的搜索结果中会包含其他的新闻平台的内容，检索结果更加全面；二是和百度、谷歌等搜索引擎相比又能得到更加纯粹的新闻报道。

2.2 数据预处理

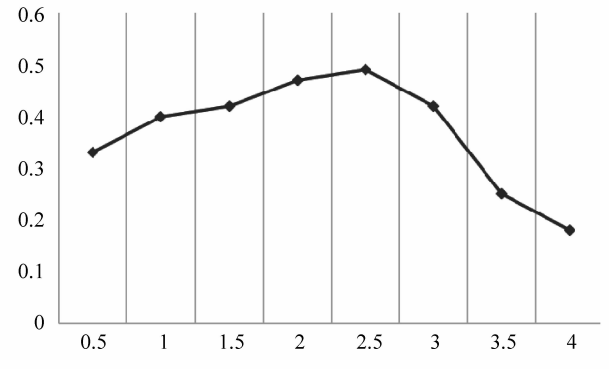
数据预处理的过程，主要包括文本数据结构化、分词、去停用词等步骤。由于抓取的新闻文档是非结构化的纯文本形式，需要将其结构化为包含时间、



(a) 常数型



(b) 线性型



(c) 指数型

图 4　不同参数下常数型、线性型、指数型衰减函数 F 值对比

位置等属性信息的标题、段落、句子。同时网页新闻的开头通常会包含报道来源等信息，在进行文档预处理时需要删除掉这些和新闻内容无关的信息，避免在进行语义分析时产生影响。

词向量的预训练方法采用 Python 语言中的 gensim 工具包进行训练，由于所使用的训练语料对模型的训练结果影响较大，这里选用的是中文维基百科的语料库（800MB）和搜狗中文新闻语料库（1.2GB），使用的语言模型是 CBOW 模型，使用 5 个词语构成的窗口，构建 128 维词向量。最后经过训练得到每个词语对应的 128 维向量表示，例如，词语“地铁”的向量表示为：（0.670 164，0.562 339，…，0.734 66）。

2.3　句子权重及权重系数计算

在句子权重计算过程中，根据预处理后的结构化文本，计算每个句子的五部分权重，各部分权重系数由两部分组成：经验系数 α 和归一化系数 ε，如式（11）所示。

$$\lambda = \alpha \times \varepsilon \tag{11}$$

各个权值的归一化系数 ε 的计算方法为：首先计算整个文档集中该权重的平均值，则归一化系数 ε 为平均值的倒数，这样可以使得经验系数 α 是与权重取值范围和分布无关的系数，同时可以通过经验系数 α 看出各个权重的重要程度。

由于句子权重包括五种权值及权重系数，不能通过单一的 F 值来优化各个系数，这里采用类梯度下降的方法，人工优化确定各个系数。以 TF-IDF 权重值系数的确定为例：

（1）首先取 TF-IDF 的权重系数为默认值 1。

（2）根据此系数，生成机器摘要。

（3）计算机器摘要所有句子的 TF-IDF 权重平均值 A 和人工摘要的所有句子的 TF-IDF 权重平均值 B。如果 A 小于 B 则调大权重系数，反之则调小。

（4）重复（2）～（3）步骤。

同理可以调整优化其余参数的取值。最终得到各个权值的系数。

2.4　句子选择及摘要生成

摘要的核心是要从原文句子中选一个句子集合，使得该集合在相关性与多样性的评测标准下得分最高，在句子选择的过程中，就需要避免选择包含重复信息过多的句子，这里采用 MMR 的方法，如式（12）所示。

$$\text{score}_{\text{MMR}}(s_i) = \underset{s_i \in D \cdot S}{\text{argmax}}\left[\lambda w_{s_i} - (1-\lambda)\max_{s_j \in S}\text{sim}(s_i, s_j)\right] \tag{12}$$

其中，s_i 表示第 i 个句子，w_{s_i} 代表的是 s_i 的权重，而 $\text{sim}(s_i, s_j)$ 代表的是冗余性，通过不断迭代计算，每次选出一个最优的句子。具体计算逻辑如下：

```
calculate similarity between each sentence
save the similarity value
for each sorted sentence
calculate MMR
select sentence of the max MMR into summary
if length of summary enough
    stop
```

```
else
    continue
```

在句子排序上，根据每个句子所在文档和在该文档中的顺序，将已选为摘要句的句子按照同一文档、出现的先后顺序进行排列。为了增加用户的可读性，同一文档的句子组成一个段落，同时根据文档时间对段落进行倒序排列。最终摘要结果示例如下：

2017-12-28 从 26 日起，武汉地铁各线路、各车站开通 N F C 安卓手机过闸和手机扫码购票服务。这意味着乘客在未携带现金的情况下，可通过移动支付乘坐地铁。武汉地铁集团首席专家朱东飞介绍，此次受惠的是武汉地铁全线网，即轨道交通 1 号线、2 号线、3 号线、4 号线、6 号线、8 号线、机场线和阳逻线共 8 条运营线路，覆盖全部 1 6 7 个车站所有闸机和售票机。

2017-12-27 通过手机下载厦门地铁官方 App 应用软件(App 名称:厦门地铁 AMTR)，即可在手机上购买地铁电子单程车票。

2017-12-22 北京地铁官方网站消息，2017 年 12 月 23 日起，北京轨道交通全路网实现线上购票、车站取票。即乘客通过北京轨道交通单程票互联网票务服务平台 App 进行线上购票，可在全路网各车站 FAM（网络取票机）上进行取票、进站乘车。

2017-12-08 12 月 7 日，记者从合肥轨道交通公司获悉，为方便乘客乘坐轨道交通，合肥轨道交通推出聚合二维码线上支付。从 12 月 5 日起，在合肥轨道交通 1 号线各站点客服中心、临时售票岗亭内购买计次票、储值票、纪念票，或者为储值票充值，均可通过扫码支付。12 月 6 日，轨道交通 2 号线百万市民大型试乘体验活动启动当天，聚合二维码支付同步上线。聚合二维码聚合微信、支付宝等支付通道，实现一码收款。此外，合肥轨道交通已启动在自助购票机上通过扫码支付购买单程票相关工作，待自助购票机二维码支付上线后，乘客乘坐轨道交通不需要携带现金，带上一部手机即可。

2017-09-04 北京青年报记者从市交通委获悉，继地铁机场线试点手机购票后，昨天 20 座大客流地铁站也可以手机购买单程票了。此外，今明两年北京地铁还将实现手机扫码进出站，预计到 2018 年一季度，北京市地铁全网所有线路和车站将实现刷手机二维码进出站。

2017-08-18 8 月 18 日，据市交通委消息，北京轨道交通单程票互联网票务的官方服务平台——易通行 APP 正式上线，率先在机场线试点运行。到 9 月 20 日，乘客只要再开通 APP 上的二维码刷闸功能，就能够直接刷二维码进出站，乘客将彻底告别现场买票排队，坐地铁会更加方便快捷。

2.5 实验结果分析

为了保证实验结果的稳定性，对于评测数据中的同一份文档集摘要，分别由三名专家在句子维度上独立标注出人工摘要，表 1 是评测数据的统计信息。

表 1 评测数据概况

查询条件	总文档数	总句子数
雄安新区规划	138	3 520
共享单车的未来发展	59	1 864
个税改革	184	4 933
宫颈癌疫苗接种	28	840
区块链技术的应用	67	1 677
阿尔法狗与柯洁	30	846
复兴号正式投入运营	12	358

由于人工摘要是在句子维度上生成的，通过对每个句子进行编号，人工标注数据集则可用相应数字表示，表 2 为各个数据样本下，对三份人工摘要进行 Pearson 相关性检验结果，其中 S1、S2、S3 分别表示三位专家的摘要。

表 2 评测数据相关性检验结果

查询条件对应数据样本集	(S1,S2)	(S2,S3)	(S1,S3)
雄安新区规划	0.766	0.716	0.754
共享单车的未来发展	0.828	0.856	0.704
个税改革	0.615	0.663	0.706
宫颈癌疫苗接种	0.763	0.843	0.726
区块链技术的应用	0.785	0.724	0.763
阿尔法狗与柯洁	0.845	0.863	0.812
复兴号正式投入运营	0.826	0.817	0.834

由表 2 可以看出，三位专家在各个数据样本上

的相关性均大于 0.6，具有较强的相关性，对于句子数量较多的样本，由于数量的影响，相关性较其他数据样本略微低。

目前对自动文本摘要的评价方法主要有两种：内部评价法和外部评价法。其中内部评价法是比较客观的，将系统生成的自动文本摘要与专家摘要采用一定的方法进行比较是目前常见的文摘评价模式。

摘要质量的评价方法采用自动摘要领域广泛使用的 ROUGE 指标，ROUGE 是一种基于召回率的自动评价方法，通过比较自动文摘中包含的基本语义单元数目在专家文摘中的数目多少来衡量[21]。ROUGE 有五种不同的评价指标：ROUGE-N，ROUGE-L，ROUGE-W，ROUGE-S，ROUGE-SU。根据本文的研究内容和文本特点，选择 ROUGE-2、ROUGE-3、ROUGE-L、ROUGE-SU4 四种具体的评价指标。

在评价过程中，对比的方法有：(1) ST-SUM：本文方法生成的结果；(2) TF-IDF：基于 TF-IDF 计算句子权重，选择摘要句的方式；(3) TextRank：基于图模型 TextRank 的方法[22]；(4) LDA：基于 LDA 主题模型，使用 KL 散度计算句子相似度的方法；(5) NY-SUM：不使用语义相似度；(6) NQ-SUM：不使用与查询句相似度权重；(7) NT-SUM：不使用时序权重系数生成的系统摘要。

在每个评测样本上，对每种摘要方法分别计算相应的 ROUGE 得分，图 5 为每种摘要方法的 ROUGE 得分在三份人工摘要上的平均值。

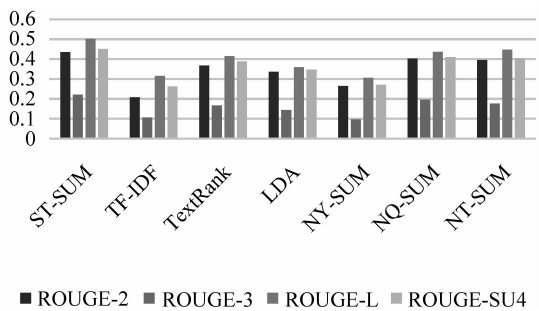


图 5　实验评测结果

为了验证本文方法与其他方法具有显著性差异，在 ROUGE 每个指标上分别进行各摘要方法与本文方法的显著性检验。表 3 为检验结果的 P 值。

表 3　各摘要方法与本文方法的显著性检验结果

摘要方法	ROUGE-2	ROUGE-3	ROUGE-L	ROUGE-SU4
TF-IDF	0.005 2	0.001 3	0.004 3	0.006 7
TextRank	0.009 1	0.010 2	0.013 4	0.015 6
LDA	0.008 6	0.007 5	0.009 3	0.012 4
NY-SUM	0.001	0.002 6	0.004 7	0.001 6
NQ-SUM	0.015	0.008 1	0.021	0.012
NT-SUM	0.009 9	0.008 7	0.01	0.021

从以上结果的分析可以看出：基于本文的方法在 ROUGE-2、ROUGE-3、ROUGE-L、ROUGE-SU4 指标上均要高于其他方法，得分稳定性较高；从显著性检验结果可以看出本文方法与其他几种方法具有显著性差异，说明该方法得到的摘要与人工摘要的标准更加接近，质量更好。通过与 NT-SUM 方法的对比可以看出，本文提出的基于时序特征的权重衰减系数对于提高摘要质量是有效、可行的；通过与 NQ-SUM 结果对比可以看出，本文方法使用的基于查询的相似度权重计算方法，得到的结果更能符合用户的查询需求；通过与 NY-SUM、LDA 的对比可以看出，不使用相似度计算的 NY-SUM 方法得分要低于本文方法和基于 LDA 的方法得分，同时基于 LDA 的得分仍低于本文方法，可见在文档集内容相似程度比较高的情况下，使用基于 Word2Vec 的空间余弦相似度计算方法仍有较好效果。

3　结语

本文针对用户日常阅读需求较高的新闻文本，提出了一种基于查询的新闻多文档自动摘要方法，该方法同时考虑了新闻文本的倒金字塔结构、时效性等特点，对相似度计算、句子权重值分布等进行改进，使得摘要质量得到提升。同时结合用户在阅读新闻过程中对最新事件的关注度更高的特点，对句子权重增加了时间序列上的衰减系数，使得距今时间越近的句子权重越高。通过实验对比分析，本文提出的方法在人工摘要上的评分和用户调研打分上都有较好的表现。但本文仍存在一些不足，在相似度的计算中，没有考虑语义相反的情况，对于语义相反的句子，其相似度会高于其他类型的句子。在查询意图的分析上，当用户输入查询语句查询时，应分析用户真正的查询意图，以使检索得到的文章、查询语句与句子相似度的计算等部分更加准确。

参考文献

[1] Cai X, Li W. Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization[J]. IEEE Transactions on Audio Speech & Language Processing, 2012, 20(5): 1597-1607.

[2] Canhasi E. Query focused multi-document summarization based on five-layered graph and universal paraphrastic embeddings[C]// Proceedings of Artificial Intelligence Trends in Intelligent Systems, 2017: 220-228.

[3] Xiong S, Ji D. Query-focused multi-document summarization using hypergraph-based ranking[J]. Information Processing & Management, 2016, 52(4): 670-681.

[4] Zheng H T, Guo J M, Jiang Y, et al. Query-focused multi-document summarization based on concept importance [M]. Advances in Knowledge Discovery and Data Mining, 2016: 443-453.

[5] 徐晓丹. 基于子主题和用户查询的多文档摘要系统[J]. 计算机系统应用, 2011, 20(3):112-115.

[6] Luo W, Zhuang F, He Q, et al. Exploiting relevance, coverage, and novelty for query-focused multi-document summarization[J]. Knowledge-Based Systems, 2013, 46(1): 33-42.

[7] Yang L, Cai X. Semi-supervised co-clustering for query-oriented theme-based summarization[J]. Research Journal of Applied Sciences Engineering & Technology, 2012, 4(18): 3410-3414.

[8] Naveen G K R, Nedungadi P. Query-based multi-document summarization by clustering of documents [C]// Proceedings of International Conference on Interdisciplinary Advances in Applied Computing, 2014(58).

[9] Yang G. A novel contextual topic model for query-focused multi-document summarization[C]// Proceedings of IEEE International Conference on TOOLS with Artificial Intelligence, 2014: 576-583.

[10] Sun R, Wang Z, Ren Y, et al. Query-based multi-document abstractive summarization via submodular maximization using event guidance[M]. Web-Age Information Management, 2016: 310-322.

[11] Li J, Li S. Query-focused multi-document summarization: Combining a novel topic model with graph-based semi-supervised Learning [J]. arXiv: 1212.2036. 2012.

[12] Shen C, Li T. Learning to rank for query-ocused multi-document summarization [C]//Proceedings of IEEE International Conference on Data Mining, 2012: 626-634.

[13] Feigenblat G, Roitman H, Boni O, et al. Unsupervised Query-focused Multi-document summarization using the cross entropy method [C]//Proceedings of the International ACM SIGIR Conference, 2017: 961-964.

[14] Nema P, Khapra M, Laha A, et al. Diversity driven attention model for query-based abstractive summarization[C]//Proceedings of Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1063-1072.

[15] Liu Y, Zhong S H, Li W. Query-oriented multi-document summarization via unsupervised deep learning[J]. Expert Systems with Applications, 2012, 2(21):35-47.

[16] Rush A M, Chopra S, Weston J. A neural attention model for abstractive sentence summarization [J]. arXiv:1509.00685. 2015.

[17] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [J]. arXiv:1301.3781. 2013.

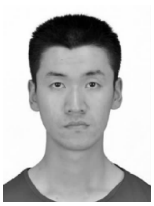
[18] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries[C]// Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998: 456-459.

[19] Valizadeh M, Brazdil P. Exploring actor-object relationships for query-focused multi-document summarization[J]. Soft Computing, 2014, 19(11): 1-13.

[20] 王飞, 谭新. 一种基于 Word2Vec 的训练效果优化策略研究[J]. 计算机应用与软件, 2018(1):97-102.

[21] Lin C Y, Och F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics, 2004:605-612.

[22] 李娜娜, 刘培玉, 刘文锋, 等. 基于 TextRank 的自动摘要优化算法[J]. 计算机应用研究, 2019(5):1-3.



王凯祥(1994—),通信作者,硕士研究生,主要研究领域为自然语言处理、自动文本摘要。
E-mail:wkx@ruc.edu.cn

任明(1980—),博士,副教授,主要研究领域为商务智能与竞争情报、大数据分析。
E-mail: renm@ruc.edu.cn