doi:10.16018/j.cnki.cn32-1650/n.202101007

基于事件时序关系的自动摘要抽取

陈　红
(安徽理工大学 计算机科学与工程学院,安徽 淮南　232001)

摘要:由于文本中事件之间的时序关系可以帮助人们更好地理解文本内容,故针对新闻报道类文本,将事件作为其基本语义单元,并根据时序关系建立事件有向网络文本表示模型;利用 PageRank 算法结合主题相关度对时序网络进行节点重要度计算及调整;最后,按照重要度以及事件发生的顺序进行排序,并按照一定的压缩比提取摘要句,删除冗余的句子,将事件对应的原语句作为摘要。实验结果表明,基于事件时序关系的自动摘要方法效果较好。

关键词:时序关系;时序网络;自动摘要;PageRank 算法
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:1671-5322(2021)01-0031-05

移动互联网时代的到来,海量文本信息呈现在人们面前。为了更快更好地从这些文本中获取有价值的信息,自动摘要提取技术尤为重要。文本摘要提取技术主要分为两类:一类是生成式的自动摘要提取技术,另一类是抽取式的自动摘要提取技术。

生成式的技术是通过计算机对文本内容进行全面深刻地理解,生成语法正确、逻辑合理能概括文本内容的摘要句子。此类方法过度依赖计算机的自然语言理解能力以及严谨的句子生成技术,运用起来过于复杂。而抽取式的摘要技术是通过对文本多方面的信息进行分析,提取原文句子进行组合成为摘要句。该方法较为灵活通用,成为自动摘要领域研究方法的主要技术。

1　相关工作

自动摘要提取技术的概念最早由 Luhn[1] 提出。自动摘要抽取的方法最先开始于统计的方法,因为基于统计的方法相对于其他方法,模型简单,实现起来也比较快。因此,在统计思想的基础上,Lloret 等[2] 提出了基于统计模型的自动摘要方法,即通过混合概率模型建立特征词的权重值和识别词之间的语义关系,然后将提取的特征词对应的句子按权重排序,并提取出来确定摘要。Lynn 等[3] 针对统计特征中词频不够全面的问题,提出将句子本身的其他特征与通用统计特征相结合以此来提高权重计算的精度。

随着自动摘要抽取研究的深入,基于图模型的自动摘要抽取技术被提出。Erkan 等[4] 提出一种基于图模型的文本摘要提取方法,将句子集作为文本的单元,构造一个以句子为顶点的图,并用图来计算句子的重要性,再根据重要性排名来提取句子并组合成摘要。张云纯等[5] 提出一种基于图模型的多文档摘要生成算法,对海量文本进行主题划分,在特征向量化和主题划分方面对传统的方法进行了改进,并为新闻文本的节点重要度计算设计了相应的数学计算公式。

基于图模型的摘要抽取技术的迭代计算特点能较好地获取文本中全局的特征信息,获得较为精确的句子权重,最终提高了抽取出来的摘要质量。但传统的图模型大多通过对文本的各项特征进行提取,将重心放在节点之间的关联度并对句子权重进行评估,忽略了文本中事件发生的逻辑顺序的作用,并且多以句子为基本单元进行图模型的构建。针对这一问题,本文提出了一种基于事件时序关系的自动摘要抽取方法,即以事件为基本语义单元,结合已标注的文本中抽取出的事件时序集合,建立事件之间的时序网络有向图(这里的时序关系集合是广义上的时序,包括因果、递进、并列、伴随等事件之间的非分类关系都

收稿日期:2020-07-19
作者简介:陈红(1995—),女,安徽池州人,硕士生,主要研究方向为智能信息处理。

被归纳于时序关系中,因此构建出的网络图能够很好地表现出文本各个事件之间的关系);再通过事件的相似度计算来合并高相似度的事件节点,从而减少摘要抽取过程中的冗余度和重复的工作量;在此基础上结合 PageRank 算法来计算节点重要度,并进行适当的权重调整;最终按照节点重要度和事件发生的先后顺序进行摘要句排序,并按照一定的压缩比提取出摘要句。

2 事件和 CEC 语料库

2.1 相关定义

定义 1 事件[6]指在某个特定的时间和环境下发生的、由若干角色参与,并表现出若干动作特征和影响的一件事情。

定义 2 事件时序关系 时序关系分为事件与时间之间的关系、事件与事件之间的关系,本文研究的是文本中任意事件对之间的时序关系,即给定事件对$\langle e_i, e_j\rangle$,能够得到事件对之间的时间关系 r。其中事件对集合 $C_{EE}=\{e_1, e_2, \cdots, e_n\}$,事件时序关系集合 $R=\{$before, after, overlap, unknown$\}$。

定义 3 事件相似度[7]指事件的相似程度,通常用区间[0, 1]之间的值来表示。对于任意事件 e_i 和 e_j 的相似度,利用事件要素的相似度进行计算,即根据现有语料中易获取的对象、动作、环境 3 个要素来计算事件相似度,公式如下:

$$\mathrm{sim}(e_i - e_j) = \sum_{k=1}^{3} w_k s(e_{ik}, e_{jk})$$

式中:e_{ik}和 e_{jk}分别表示事件 e_i 和 e_j 的第 k 个要素;w_k 表示事件各要素在计算相似度时的权重,区间为[0,1]。

通过对事件要素对应的词汇结合同义词林来判断事件间对应要素的相似性。根据事件要素在文本中的描述能力将对象、动作、环境 3 个要素的权重分别定义为:$w_1=0.5, w_2=0.3, w_3=0.2$。通过实验观察,当事件相似度 $\mathrm{sim}(e_i, e_j)\geqslant 0.7$ 时,认为两事件相似。

定义 4 事件时序网络 一组包含一系列事件节点及相连边的有向图的集合。用结点 V 表示事件、边 E 表示事件间的时序关系,则事件时序网络可表示为 $G=[V, E]$。

2.2 CEC 语料库

CEC(Chinese Emergency Corpus)语料库主要是对事件及其事件要素的标注,是从各大新闻网站收集关于地震、交通事故、火灾、恐慌袭击以及食物中毒 5 类国内外突发事件的中文新闻报道[8],是专门针对突发公共事件的中文语料库。本文采用 CEC 2.0 语料库(由上海大学语义智能实验室构建),CEC 2.0 是在 1.0 版本上对文本进行了扩充,统计结果如表 1 所示。

表 1 CEC 各项统计结果

Table 1 Statistical results of CEC

新闻报道类型	篇数	事件数量
地震类	62	1 001
火灾类	75	1 216
交通事故类	85	1 790
恐怖袭击类	49	823
食物中毒类	61	1 109
总计	332	5 991

3 基于事件时序关系的自动摘要抽取

3.1 事件时序网络的构建

构建事件时序网络,首先从 CEC 语料库已标注的文本中任选一篇文本 D,从文本 D 中抽取已经标注好的事件元素,得到事件集合 $E=\{e_1, e_2, e_3, \cdots, e_n\}$和事件时序关系集合 R。在此基础上构建事件网络,步骤如下:

Step 1 分别对节点集合 V_D、有向边集合 E_D 进行初始化。

Step 2 在事件集合 E 中依次提取事件,并一一映射到事件时序网络图中的节点,最终得到事件节点集合 $V_D=\{v_1, v_2, \cdots, v_i, v_j, \cdots, v_k\}$。

Step 3 从事件节点集合 V_D 中选取节点 v_i 作为事件时序网络中的任意节点,在时序关系集合 R 中依次查找与 v_i 有时序关系的节点 v_j;如果节点 v_i 和 v_j 间具有 before 和 after 时序关系的添加一条有向边 e_{ij},时序关系为 overlap 的添加双向有向边 e_{ij},时序关系为 unknown 的不添加边。

Step 4 从事件节点集合 V_D 中任取节点 v_i,再一一遍历集合 V_D 中的其他事件节点 v_j,计算它们对应的事件 e_i 和 e_j 的相似度。若相似度为 1,则将两个节点合并;如果相似度大于或等于阈值但是小于 1(阈值设定为 0.7),则在 v_i 和 v_j 之间添加双向的有向边;若相似度小于阈值,则不添加边。

Step 5 根据步骤 3、步骤 4,可以得到有向图集合 $E'_D=\{e_{12}\cdots, e_{ij}\cdots\}$,从而得到文本 D 的事件时序网络有向图。

Step 6 通过以上步骤建立的时序关系网络较为复杂,且存在多个时序回路。为了更好地表示文本中事件发生、发展的过程,避免时序网络中出现回路,保证网络的单向连通性,对事件时序关系网络进行简化,简化规则如图 1 所示。图 1 中,若一个事件 e_1 同时指向两个事件 e_2 和 e_3,并且另外两个事件 e_2 和 e_3 之间也存在单向边,此时事件 e_3 与 e_1 之间就形成了一个回路,则将 e_1 与 e_3 之间的边去除。

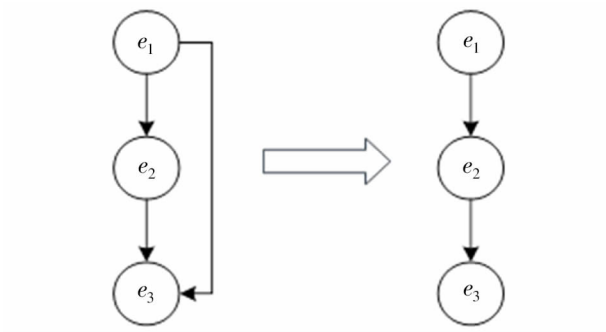

图 1　时序关系网络简化规则
Fig 1　Simplification rules for temporal relationship networks

为理解构建事件网络步骤的全过程,用网络事件实例进行说明。选取 CEC 语料库中“北京北六环 4 车连撞拥堵 10 公里”的新闻文本 D(包含 5 个段落、11 个句子),利用以上步骤方法得到文本 D 的事件时序网络有向图(包含 32 个节点、34 条有向边),再通过可视化软件 NetDraw 将其可视化,得到图 2 所示的事件时序关系网络。图 2 中每个节点使用事件及其动作要素进行标识,节点之间带箭头的线表示事件之间的时序关系。

3.2　自动摘要生成

自动摘要的抽取,是将文本中概括出文本大概信息的句子按照一定的顺序排列组合形成的片段。通常文本在对事件进行描述时会有一定的表达规律,若整个文本描述的是一个类型的事件,那么该文本会以相似的语言对同一事件类型进行描述。这类相似的语言表达,可以进行合并,以减少摘要抽取过程中内容的重复率,提高摘要的精度。本文所采用的 CEC 语料就是一个按照事件类型划分的新闻类文本,因此在图模型构建时,不仅用到了事件之间的时序关系,还对事件节点之间的相似度进行了计算。

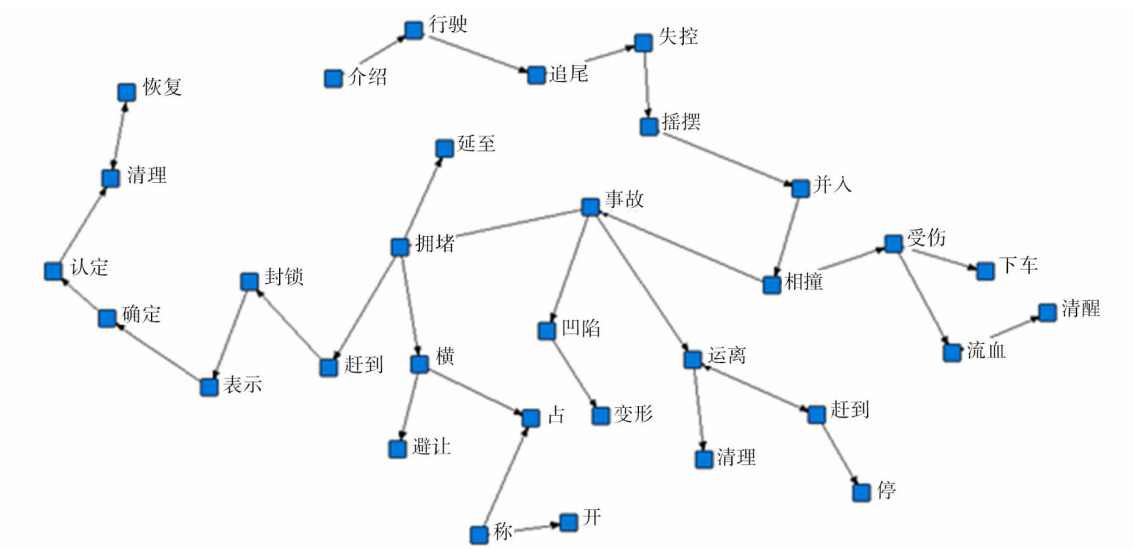

图 2　事件时序关系网络
Fig 2　Event temporal relationship network

根据事件之间的时序关系构建的有向图模型,既可以表达出事件的发展趋势,又能表示出事件之间的关联性。如果一个事件与其他事件的关联太少,那么这个事件节点可以被认为是不重要的,则该事件对摘要抽取的贡献不大,可以不被考虑为摘要句。

因此对于事件节点的重要度计算不仅要考虑自身的重要度,还要考虑和其他节点间的重要度。由于抽取时序关系的对象是一篇文本中的任意事件对,且从图 2 的时序网络图中可以发现,各节点之间都存在一条甚至多条有向边,故可以将问题转化为链接矩阵来表示节点的链入和链出。现采

用 PageRank 算法对节点的重要度进行排序,事件节点 e_i 的重要度计算公式为:

$$S(e_i) = (1-d)/n + d\sum_{i\in In(e_i)}\frac{S(e_j)}{Out(e_j)}$$

式中:e_i 和 e_j 是事件图中的任意节点;$S(e_i)$ 表示事件 e_i 的重要度;$In(e_i)$ 表示连接线指向 e_i 的集合(导致事件 e_i 出现的事件总数);$Out(e_j)$ 表示连接线指向别的事件 e_k 的集合(由 e_j 导致 e_k 出现的事件总数);n 为图中的节点数;d 为阻尼系数,通常取 0.85。

在 CEC 语料中,每篇新闻文本都有一个相应的标题,这在一定程度上反映了文本的主旨内容,因此在正文文本中与标题相似度较高的句子就有更大的可能性成为摘要句。所以在计算时序网络图中各个节点的重要度后,需对每个节点对应的句子与文本标题进行相似度计算。

现使用 Doc2Vec 对句子进行向量化,利用余弦相似度衡量文中句子与标题的相似性。若句子与标题具有较高的相似性,则对该句子的最终权重进行调整,调整规则如下:

$$S(e_i) = S(e_i) + \mathrm{sim}(s_i), 0.7 < \mathrm{sim}(s_i) \leqslant 1$$

$$S(e_i) = S(e_i), 0 \leqslant \mathrm{sim}(s_i) \leqslant 0.7$$

式中:$S(e_i)$ 为节点的最终权值,$\mathrm{sim}(e_i)$ 为节点 e_i 所在句子 s_i 与文档标题的余弦相似度。

将调整后的最终权值作为时序网络有向图中节点的重要度,结合该节点事件在时序关系集合中的时序关系进行综合排序;取文本的摘要压缩比为 30%,将多余的节点删除;最后根据各事件节点所在的句子进行提取,最终完成摘要的抽取工作。

4 实验结果与分析

将 CEC 语料库作为实验语料,从语料库的 5 个事件类中各随机抽取 10 篇,共计 50 篇文本语料,采用内部评测法,分别以摘要准确率 P、召回率 R 和 F 值对摘要质量,即将通过自动摘要抽取的摘要句子与人工抽取的摘要句子进行比较、评估。其中 F 值是摘要准确率和摘要召回率的调和平均值。

$$P = \frac{\text{准备提取的主题句数目}}{\text{所有提取的主题句数目}} \times 100\%$$

$$R = \frac{\text{准备提取的主题句数目}}{\text{所有人工标注的主题句数目}} \times 100\%$$

$$F = \frac{2PR}{P+R} \times 100\%$$

为了验证自动摘要方法的有效性,选取其他文档自动摘要方法同本文方法进行实验对比。对比结果如表 2 所示。

表 2 各方法实验效果

Table 2 Experimental results of each method

方法	实验语料数据	实验结果		
		P	R	F
结合 Bigram 语义扩充的事件摘要法[9]	2017 年某电信公司业务部门一个月的投诉工单,共 2 256 条文本数据	0.22	0.86	0.34
基于事件要素的自动文摘抽取[10]	从 CEC 语料库的 5 类事件中分别随机抽取 8 篇,共 40 篇文本语料	0.66	0.59	0.62
基于主题聚类的多文本自动摘要算法[11]	从搜狗实验室新闻数据集中选取 4 个主题事件集合作为实验语料	0.69	0.54	0.61
本文方法	从 CEC 语料库的 5 个事件类中各随机抽取 10 篇,共 50 篇文本语料	0.79	0.62	0.69

从表 2 可以看出,本文所用方法在 P、R 和 F 值上都取得了不错的效果。虽然各方法所用语料不太相同,但是在针对新闻类文本的摘要抽取效果方面,本文方法在各个评分上也都优于其他方法,这进一步验证了本文方法的有效性。

5 总结

由于事件之间的时序关系可以很好地表达事件之间的关联性,以及逻辑顺序,方便人们更好地理解文本中事件的发展脉络。因此,针对突发事件的新闻类文本,提出了一种基于事件时序关系的文本摘要抽取方法,即根据事件之间的时序关系,以事件为节点,以事件间的时序关系并结合事件相似度为边,构建事件时序有向图模型;根据构建的时序网络图模型,结合经典的 PageRank 算法对图中节点进行重要度计算,再根据事件所在句

子与文本主题句的相似度进行最终的权重调整;最后,对事件节点对应的句子按照权重大小以及事件发展先后顺序进行排序,并按照一定的压缩比提取摘要句,删除冗余的句子,从而得到最终的摘要句子。通过实验证明了本文方法抽取的摘要效果较好。

参考文献:

[1] LUHN H P. The automatic creation of literature abstracts[J]. IBM Journal of Research and Development, 1958,2(2):159-165.

[2] LLORET E, PALOMAR M. Text summarisation in progress:a literature review[J]. Artificial Intelligence Review, 2012,37(1):1-41.

[3] LYNN H M, CHOI C, KIM P. An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms[J]. Soft Computing, 2018,22(12):4013-4023.

[4] ERKAN G, RADEV D R. LexRank:Graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research 2004,22(1):457-479.

[5] 张云纯,张琨,徐济铭,等. 基于图模型的多文档摘要生成算法[J]. 计算机工程与应用,2020,56(16):124-131.

[6] 廖涛,付维成,方贤进. 基于正负加权的中文事件识别研究[J]. 计算机应用与软件,2019,36(11):175-181,217.

[7] 杨竣辉,刘宗田,刘炜,等. 基于文本事件网络自动摘要的抽取方法[J]. 计算机科学,2015,42(3):210-213,223.

[8] 刘炜,王旭,张雨嘉,等. 一种面向突发事件的文本语料自动标注方法[J]. 中文信息学报,2017,31(2):76-85.

[9] 吴佳伟,曹斌,范菁,等. 一种结合 Bigram 语义扩充的事件摘要方法[J]. 小型微型计算机系统,2019,40(7):1380-1385.

[10] 孙佩佩,廖涛,刘宗田. 基于事件要素的自动文摘抽取[J]. 计算机与数字工程,2015,43(10):1829-1833.

[11] 徐小龙,杨春春. 一种基于主题聚类的多文本自动摘要算法[J]. 南京邮电大学学报(自然科学版),2018,38(5):70-78.

Automatic Summary Extraction Based on Event Temporal Relationship

CHEN Hong

(School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan Anhui 232001, China)

Abstract:Since the temporal relationship between events in the text can help people better understand the content of the text, for news report texts, events are used as the basic semantic unit, and a directed network text representation model of events is established according to the temporal relationship. PageRank algorithm combined with topic relevance is used to calculate and adjust node importance of time series network. Finally, the sequence is sorted according to the importance and the sequence of events, and the abstract sentence is extracted according to a certain compression ratio, and the redundant sentence is deleted, and the original sentence corresponding to the event is taken as the summary. The experimental results show that the automatic summarization method based on event sequence relation is effective.

Keywords:sequential relation; temporal network; automatic summary; PageRank algorithm

(责任编辑:李华云)