·人工智能及识别技术· 文章编号:1000-3428(2017)08-0219-06 文献标志码:A 中图分类号:TP311

基于加权 TextRank 的新闻关键事件主题句提取

蒲 梅,周 枫,周晶晶,严 馨,周兰江

(昆明理工大学 信息工程与自动化学院,昆明 650500)

摘 要: 为了在大量的新闻中快速找到自己感兴趣的内容,提出在单文档中基于加权 TextRank 算法提取主题句的方法,以得到新闻关键事件信息。通过计算新闻文本句子关键词的互信息值,对新闻报道进行事件句和非事件句的分类,过滤出非事件句。基于 TextRank 算法的思想,构建一个事件句有向图,引入句子位置、句子相似度和关键词覆盖频率 3 个影响因子,以此计算句子之间的影响权重,利用 TextRank 模型对图中的每个点计算权重,并选取排序最靠前的句子作为关键事件的主题句。实验结果表明,该方法的抽取效果优于基于词频-逆文档概率和新闻标题的主题句抽取方法。

关键词: TextRank 算法;句子相似度;关键事件;主题句提取;影响权重

中文引用格式: 蒲 梅,周 枫,周晶晶,等. 基于加权 TextRank 的新闻关键事件主题句提取[J]. 计算机工程,2017,43(8):219-224.

英文引用格式: Pu Mei, Zhou Feng, Zhou Jingjing, et al. Topic Sentence Extraction of Key News Events Based on Weighted TextRank[J]. Computer Engineering,2017,43(8):219-224.

Topic Sentence Extraction of Key News Events Based on Weighted TextRank

PU Mei,ZHOU Feng,ZHOU Jingjing,YAN Xin,ZHOU Lanjiang

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

【Abstract】 In order to quickly find the content you are interested in in large number of news, a method based on weighted TextRank algorithm is proposed to extract the topic sentence in a single document and get information about key news events. It classifies news reports as event sentences and non-event sentences and filters the latter by calculating the mutual information value of the keywords in the news text sentences. It constructs a directed graph of event sentences on the basis of TextRank algorithm, and calculates the influence weight between sentences by introducing three influence factors of the sentence position, sentence similarity and keyword coverage frequency. It calculates the weight for each point in the graph by using TextRank model and selects the most front sorting sentences as topic sentences of the key events. Experimental results show that the proposed method is better than the methods based on Term Frequency-Inverse Document Probabilistic(TF-IDF) and news title in topic sentence extraction.

【Key words】 TextRank algorithm;sentence similarity;key event;topic sentence extraction;influence weight

DOI: 10.3969/j.issn.1000-3428.2017.08.037

0 概述

随着计算机技术和互联网的不断发展,互联网已经成为人们获取新闻信息的主要信息来源之一。然而由于互联网上新闻网页内容的急剧增加,读者很难从众多的信息中快速、准确地获取自己感兴趣的信息。含关键事件的主题句可以简洁、准确地描述出新闻报道的事件,因此利用机器学习、自然语言处理技术从多个事件中识别出关键事件及其主题句,有利于人们快捷、准确地获取事件信息。针对该问题,本文提出基于 TextRank 加权的新闻关键事件主题句提取方法。

1 相关研究

对新闻主题词、句的研究中,文献[1]为适应 Internet 时代和大规模文献处理的需要,以中文文本为处理对象,研究从主题词、主题概念和主题句 3 个不同层面自动抽取文本主题的方法,对新闻类文献做实验,并简单进行性能分析;文献[2]通过计算两个词在一个句子中的共现频率,利用每句话中的两

基金项目: 国家自然科学基金“基于篇章特征的越南语新闻事件信息抽取关键技术研究”(61562049)。

作者简介: 蒲 梅(1991—),女,硕士研究生,主研方向为数据挖掘、自然语言处理;周 枫,副教授、硕士;周晶晶,硕士研究生;严 馨、周兰江,副教授、硕士。

收稿日期: 2016-06-13 **修回日期:** 2016-08-18 **E-mail:** 731817113@qq.com

个词之间的共现频率计算其信息量,积累之后经过调整作为句子权重;文献[3]通过一些“综上所述、提出”等标引词抽取科技文献中的创新点,也可以看成一种特定文本主题句的提取;文献[4]采用多种权重度量方式,在综合评估句子反映主题价值的基础上,保证每一个主要的主题都有对应的主题句被选中,并解决主题句的去重问题,进一步提高主题句的主题覆盖性和概括性;文献[5]把主题句抽取看成分类问题,利用机器学习的方法训练一个判断主题句和非主题句的二元分类器来抽取主题句,在个人履历生成的具体应用中,取得较好的效果;文献[6]针对网页文本结构少、噪声大的特点,将句子看作点,将句子间的相似性看作边,用句子关系图描述文本中句子间的关系,抽取文本主题句的任务转换为搜索图中最多边的点;文献[7]在讨论分析 k-means,k-medoids 等聚类算法的基础上,提出一种基于聚类算法的主旨句提取方法,实验结果表明,在提高聚类准确性的基础上,新方法较其他聚类算法能够更加有效地避免遗漏主题的问题;文献[8]利用文档内句子之间的语义相关性,实现文本主题的自动生成:先对文本进行切词和分句处理实现信息分割,再结合文本聚类技术对文本句进行聚类实现信息合并,最后从每类中抽取代表句生成文本主题;文献[9]提出一种基于加权模糊聚类和互信息的主题句提取方法,使主题句尽可能全面覆盖全文主题的同时,缩减自身的冗余,以提高摘要效率。

文献[10]提出单文档层次提取包含关键事件 5w 要素的主题句思想,先假设新闻报道中有 n 个句子,且句子之间相互独立,然后综合词频、句子位置、句子长度、命名实体、句子标题重合等信息,分配对应的系数,不断地和手工结果做比较后才能得到较为合适的系数和阈值;文献[11]将文本表示为句子的线性序列,句子表示为词的线性序列,对每个句子都预处理为含有实词的词汇链,并基于知网(HowNet)计算相邻句子的相似度;文献[12]基于文档句构建无向图,将主题句的抽取问题转换为无向图中节点的权重计算问题。该方法在不同的压缩比下主题句抽取结果接近于人工摘要,召回率和准确率综合指数较高;文献[13]利用 LDA 的主题影响力计算,进而对 TextRank 算法进行改进,将候选关键词的重要性按照主题影响力和邻接关系进行非均匀传递,并构建新的概率转移矩阵用于词图迭代计算和关键词抽取;文献[14]提出基于 LDA 模型的主题句抽取方法,通过多个侧面对目标主题的衬托,采用 LDA 模型对主题信息进行建模,利用各个主题概率分布的平滑进度进行候选句的可信度计算来抽取主题句。

新闻报道是事件的载体,一篇新闻报道中可能出现多个“原子事件”,然而这些“原子事件”往往是对“关键事件”的补充。通常人们对于一篇新闻报道,更加注重的是这篇新闻报道的“关键事件”,目前从单篇新闻报道中抽取“新闻要点”的事件抽取研究,国内外学者都做了大量研究,并取得了一定效果。这些研究主要针对新闻事件的原子事件和主题事件这 2 个方面的信息抽取,但是原子事件级别的事件抽取粒度过细、实用性不足,主题级别的事件抽取粒度较粗、精确提取事件信息的效率较差,适合热点话题发现、话题分析等研究。

针对以上问题,本文以新闻事件要素分析为基础提出一种新的新闻事件主题句提取算法。该算法对新闻文本进行处理,计算文档句子关键词的 MI 值,将文档中的句子划分为事件句和非事件句,利用 TextRank 模型对事件句建立关系图,然后引入位置关系、句子相似度和关键词覆盖率 3 个影响因子,对关系图中的边添加权重,从而迭代计算出权重最高的句子作为关键事件主题句。

2 关键事件识别

2.1 基本定义

定义 1(事件 5W1H 要素)新闻报道中用来描述特定新闻事件的元素,包括主体(subject)/客体(object)、时间(time)、地点(location)、起因(cause)、谓词(predicate)、结果(result),分别对应于新闻事件要素的 5W1H(who/whom,when,where,why,what,how),用六元组 $<S/O,T,L,C,P,R>$ 表示。其中,$<S/O,P>$ 是关键要素;$<T,L,C,R>$ 是辅助要素。

定义 2(事件句)包含新闻事件 5W1H 要素 $<S/O,T,L,C,P,R>$ 中的关键要素 $<S/O,P>$ 和至少一个辅助要素 T,L,C 或 R 的句子,称为新闻事件句。

定义 3(关键事件)在新闻报道中,多个对某一事件补充说明的原子事件为关键事件。

定义 4(关键事件主题句)在新闻报道中,包含关键事件的句子称为关键事件主题句。

2.2 主题句抽取流程

2.2.1 预处理

主要利用自然语言处理(Natural Language Processing,NLP)技术对新闻文本进行处理,首先对新闻正文进行分句,并过滤新闻正文中的停用词。利用 ICTCLAS(Institute of Computing Technology,Chinese Lexical Analysis System)中文处理模块进行分词、词性标注和命名实体识别。

2.2.2 事件句识别

通过计算句子关键词的 MI 值,对新闻报道中包含事件的句子和不包含事件的句子进行分类,过滤掉非事件句。

2.2.3 TextRank 权值计算

在事件句之间构建 TextRank 模型,通过句子位

置、句子相似度、关键词出现频率对有向边进行加权,计算每个句子之间的影响力权重,根据句子之间的权重递归计算出全部事件句的权重,按照权重的大小排序输出。

2.2.4　关键事件主题句提取

在计算完权重之后,根据权重选择排名前 3 个句子作为关键事件主题句。基于 TextRank 加权算法提取新闻关键事件的主题句流程如图 1 所示。

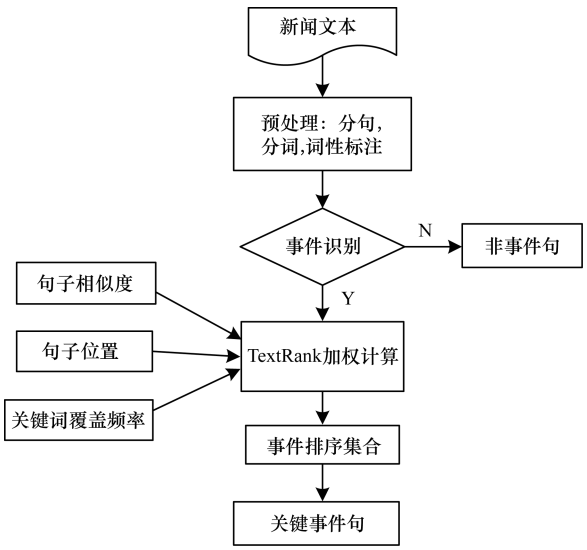

图 1　主题句抽取流程

2.3　事件句识别

新闻报道通常包含一个关键事件、多个对其补充说明的原子事件,含事件的句子称为事件句,而不包含事件信息的句子称为非事件句。为了有效抽取新闻事件的主题句,必须将非事件句过滤掉。具体步骤如下:

1)预处理

将新闻文本以句子为单位切分;并将切分后的句子进行分句、分词、词性标注、命名实体识别、停用词过滤等;其中句子的切分采用句子结束符隔开的方法进行,句子结束符包括{,;。!?}等。

2)特征提取

在步骤 1)的基础上,主要选取命名实体和关键词作为特征。关键词的提取采用词语位置加权的 TextRank 方法,该方法优于传统的 TextRank 方法和基于 LDA 主题模型的关键词抽取方法[15-16]。

3)事件识别

标题和关键词往往包含事件,但是新闻标题也有不包含新闻事件的情况,所以选择关键词作为事件句和非事件句分类的标准。

关键词在句中出现的个数越多,说明句子是事件句的可能性越大。本文对新闻文档中句子的关键词计算 MI 值,来区分句子的排名,MI 值越高则说明句子是事件句的可能性越大。此处的句子是经预处理后切分过的句子。T 表示关键词,S 表示句子,则关键词 MI 计算如下:

$$I(T,S)=\sum_{0\leqslant i\leqslant count}p(t_i,s)\,\mathrm{lb}\left[\frac{p(t_i,s)}{p(t_i)p(s)}\right] \tag{1}$$

其中,$p(t_i,s)$ 为第 i 个关键词在 s 句子中出现的频率;$p(t_i)$ 为第 i 个关键词在文本中出现的频率;$p(s)$ 为 s 句子的长度占整个文本长度的比例;$count$ 为关键词数目。MI 值为 0 时则说明句子为非事件句,若 MI 值越大,则说明句子为事件句,而且成为关键事件句的可能性越大。

3　关键事件的主题句提取

TextRank 模型源于 PageRank,通过将文本分割成若干个处理单元,对这些单元构建图模型,利用图中的节点相互投票,对图中的节点进行排名。TankRank 模型是一种基于图排序的算法,它的基本思想是“投票”,当图中的 A 节点和 B 节点有连线的时候,A 就投票给 B,节点 B 连接的节点越多,则获得的票数越多,B 点的权重就越大。TextRank 模型主要应用于关键词的提取过程,即处理单元为关键词,本文将它应用于主题句的提取中,即处理单元为切分后的句子。在关键词提取中经常将语义相同的词语进行连线,本文与关键词提取不同,认为全部的句子都是相邻的,不再提取窗口,将无向图变为加权的有向图来提取主题句。全部的句子都是相邻的原因在于:1)句子间有可能会存在影响权重大小的因子;2)这样的假设符合图模型的处理需求。

3.1　事件句图模型构建

该方法将主题句提取转化为图排序的问题,通过计算图中各个节点的权重对句子进行排名,把排名靠前的句子输出作为主题句。对事件句和非事件句进行分类后,对事件句构建有向图:

步骤 1　将文档的事件句进行句子划分,构成集合 $T=[S1,S2,\cdots,Sn]$。

步骤 2　构建事件句有向图 $G=(V,E)$,其中,V 是节点的集合,集合 T 中的句子对应于节点集合 V;E 是有向边的集合。对于一个给定的顶点 v_i,$In(v_i)$ 表示为指向该节点的点集合,$Out(v_i)$ 则为该点 v_i 指向其他点的集合,如式(2)所示。

$$W_S(v_i)=(1-d)+d\times\sum_{v_j\in In(v_i)}\frac{w_{ji}}{\sum_{v_k\in Out(v_j)}w_{jk}}W_S(v_j) \tag{2}$$

式(2)是 TextRank 递推公式,其中的 d 为阻尼系数,一般取值为 0.85;k 为图模型中的其他顶点。等式左边表示句子的权重 W_S;W_{ij} 表示图中 2 个顶点 (i,j) 之间边的权重,这里的权重加入了特征选取的 3 个影响因子,下文对 3 个影响因子做了详细描述;$W_S(v_j)$ 代表上次迭代 j 的权重,整个公式是一个迭代过程。

将 TextRank 算法的思想应用于新闻事件主题句的提取问题上,一般而言,新闻报道或者新闻片段通常由多个事件组成,这些事件中一般只有一个关键事件,其余的事件都是围绕关键事件进行描述说明,包含事件的句子称为事件句,而包含关键事件的句子称为主题句。这种关系可以通过有向图来描述,图中的点代表事件句,点之间的有向边代表它们之间存在影响权重大小的关系,边上的权重则是句子之间的影响力权重。与提取关键词不同,认为全部句子都是相邻的,不再提取窗口。这里给出了一个事件句子 v 的有向图模型,如图 2 所示。

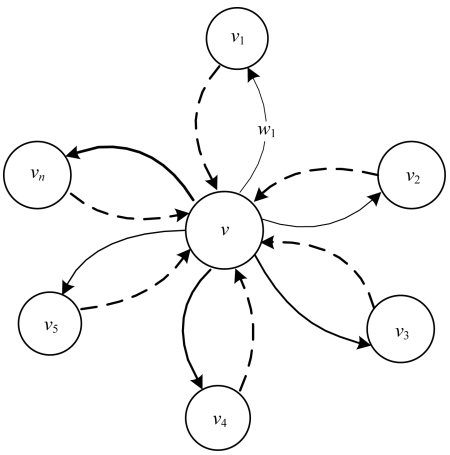

图 2 事件句有向图模型

带箭头的实线表示结点 v 指向结点 v_i 的影响力权重,虚线表示结点 v_i 指向结点 v 的影响力权重,线条的粗细表示权重大小。依据图排序算法的基本理论,本文在新闻事件识别的基础上,对候选新闻事件句之间建立“连线”构建 TextRank 模型,根据图的特点可知,图中的点代表新闻中的事件句。对文本中的所有事件句构建有向图模型形成一个全连通图。

3.2 主题句特征选取的影响因子

构建完事件句的图模型后,通过分析发现,2 个顶点之间权重的大小是影响句子排名的关键,构建 2 个句子之间的权重成为主题句提取的关键。根据提出的影响结点之间权重的 3 个影响因子,即位置关系、句子相似度和关键词覆盖率,令 W 为结点的整体权重,$\lambda_1,\lambda_2,\lambda_3$ 表示这 3 个不同的影响因子所占的比重,如式(3)所示。

$$W=\lambda_1+\lambda_2+\lambda_3=1 \tag{3}$$

针对不同的影响因子,分别给出了其计算方法。

3.2.1 位置关系

候选事件句出现的位置很重要,例如出现在首段段首、段中和段末或其他段的段首、段中和段末的位置,它的重要性随着所在的位置不同而变化。

$W_{\lambda_1}(v_i,v_j)$ 为位置关系影响因子,表示结点 v_i 的位置关系权重传递给结点 v_j 的权重,计算公式如式(4)所示。

$$W_{\lambda_1}(v_i,v_j)=\frac{L(v_j)}{\sum\limits_{v_k\in Out(v_i)}L(v_k)} \tag{4}$$

其中,$L(v_j)$ 表示结点 v_j 在文本中的位置重要性,本文根据句子所在的位置不同赋予不同的权值。对于位置的句子加权公式如式(5)所示。

$$Score_{\text{loc}}(v_j)=\frac{1}{j} \tag{5}$$

$L(v_j)$ 的取值公式如式(6)所示。

$$L(v_j)=\begin{cases}1, & v_i\\ Score_{\text{loc}}(v_j)=\dfrac{1}{j}, & v_j\end{cases} \tag{6}$$

3.2.2 句子相似度

候选事件句之间的相似度越高说明投票数越多,句子的重要性越高。

$W_{\lambda_2}(v_i,v_j)$ 为句子相似度影响因子,表示结点 v_i 的相似度权重传递到结点 v_j 的权重。计算公式如式(7)所示。

$$W_{\lambda_2}(v_i,v_j)=\frac{S(v_{jk})}{\sum\limits_{v_k\in Out(v_i)}S(v_{jk})} \tag{7}$$

其中,$S(v_{jk})$ 表示结点 v_j 所对应的句子和其他句子之间的相似度取值,其取值方式如式(8)所示。

$$S(v_{jk})=\begin{cases}\eta, & \text{2 个句子有相似度}\\ 0, & \text{2 个句子没有相似度}\end{cases} \tag{8}$$

其中,η 大于 1。用词形相似度计算 2 个句子的相似度,句子中相同的单词出现越多,说明句子的相似度越大。实验中设定的阈值 η 为 45,η 在 50 之后不同的取值对 F 值的影响比较平稳。

3.2.3 关键词覆盖率

候选事件句所覆盖的关键词越多,说明句子的重要性越高。

$W_{\lambda_3}(v_i,v_j)$ 为句子关键词覆盖率影响因子,表示结点 v_i 的覆盖率权重传递给结点 v_j 的权重。计算公式如式(9)所示。

$$W_{\lambda_3}(v_i,v_j)=\frac{F(v_j)}{\sum\limits_{v_k\in Out(v_i)}F(v_k)} \tag{9}$$

其中,$F(v_j)$ 表示节点 v_j 所对应的句子关键词覆盖率,覆盖率越高说明该节点获得的投票数就越多,对关键词覆盖率的计算就是计算句子中关键词 MI 值。

3.3 主题句提取

在上文 TextRank 公式中讲,到 W_{ij} 表示图中 2 个顶点(i,j)之间边的权重,这里可以通过式(10)计算得到。

$$W_{ij}=\lambda_1\times W_{\lambda_1}(v_j,v_i)+\lambda_2\times W_{\lambda_2}(v_j,v_i)+\lambda_3\times W_{\lambda_3}(v_j,v_i) \tag{10}$$

通过式(1)将图中各个结点 v_i 对结点 v 的影响权重进行迭代。

对于构建好的事件句图模型,已经提出了影响结点之间权重的 3 个影响因子,即位置关系、句子相

似度和关键词覆盖率。以连通图中任意一点为初始顶点,通过 TextRank 模型迭代计算图中每个节点的权重对事件句进行排名,最后输出排名前 3 名的句子作为关键事件主题句。

4　实验设置与结果分析

4.1　实验数据及评价标准

实验数据来自从新华网、人民网、搜狐网、新浪网等网站爬取的 900 篇新闻报道,内容涵盖了国际新闻、国内新闻、文化财经、科技等领域。其中,国际新闻和国内新闻的比例为 3∶2;文化、财经和科技的新闻比例为 1∶1∶1。人工标注关键事件主题句为 2 765 个,平均每篇新闻报道的关键事件主题句为 3。对这些新闻正文进行分句,并过滤掉汉语新闻正文中的停用词。中文利用 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)中文处理模块进行分词、词性标注和命名实体识别。同时通过计算关键词的 MI 值进行事件句识别,并过滤掉非事件句。

实验采用准确率 P、召回率 R 和 F 值 F 评价关键事件主题句抽取的效果,如式(11) ~ 式(13)所示。

$$P = \frac{S_{\text{correct}}}{S_{\text{extract}}} \times 100\% \tag{11}$$

$$R = \frac{S_{\text{correct}}}{S_{\text{standar}}} \times 100\% \tag{12}$$

$$F = \frac{2PR}{P+R} \times 100\% \tag{13}$$

其中,S_{correct} 是准确抽取的关键事件主题句数目;S_{extract} 是所有抽取的关键事件主题句数目;S_{standar} 是所有人工标注的关键事件主题句数目。

4.2　影响因子参数确定

为了使实验达到更好的效果,对 3 个影响因子的参数 λ_1,λ_2,λ_3 设置了不同的参数组合,其结果如表 1 所示。

表 1　3 个影响因子的参数组合

组数	λ_1	λ_2	λ_3
1	1.00	0.00	0.00
2	0.00	1.00	0.00
3	0.00	0.00	1.00
4	0.20	0.80	0.00
5	0.00	0.80	0.20
6	0.00	0.20	0.80
7	0.33	0.34	0.33
8	0.13	0.49	0.38
9	0.15	0.38	0.47
10	0.51	0.12	0.37

针对上文给出的 10 组参数组合,分别计算每篇报道的关键事件主题句抽取结果,并取其均值,实验结果如图 3 所示。

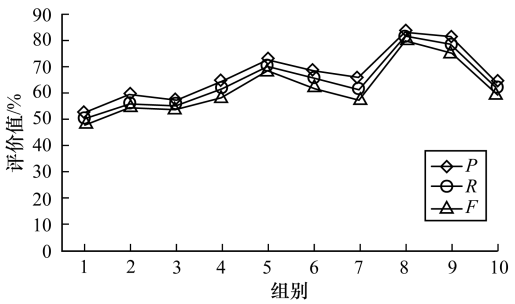

图 3　不同参数的主题句抽取结果

从表 1 和图 3 可以看出,相似度重要性和频率重要性对句子起到了重要的作用,位置重要性也对句子有影响但是影响力不大,所以参数 $\lambda_1 = 0.13$,$\lambda_2 = 0.19$,$\lambda_3 = 0.38$ 时取得的效果最好。

4.3　实验结果及分析

为了验证本文方法的有效性,采用基于词频-逆文档概率(TF-IDF)的方法[16]、基于新闻标题(Title)[10]的关键事件主题句抽取方法和本文方法(TextRank)进行对比实验,这里 3 个影响因子的参数组合为 $\lambda_1 = 0.13$,$\lambda_2 = 0.19$,$\lambda_3 = 0.38$。对 3 种方法的抽取结果进行对比,验证本文方法的性能,实验结果如表 2 所示。

表 2　不同方法下主题句抽取效果对比　　%

文章类型	评价指标	TF-IDF	Title	TextRank
国际新闻	P	75.78	78.25	81.65
	R	72.86	74.45	77.24
	F	73.60	77.18	79.93
国内新闻	P	75.34	79.61	83.36
	R	66.83	72.56	74.12
	F	71.05	75.52	77.93
文化	P	75.35	80.87	82.43
	R	71.45	74.61	76.98
	F	73.35	77.61	79.61
财经	P	73.25	79.62	81.44
	R	70.41	74.72	76.81
	F	71.81	77.09	79.06
科技	P	74.43	80.78	81.72
	R	69.61	72.82	73.14
	F	71.94	76.59	77.19

实验结果表明,本文方法在 P,R,F 3 个评价指标上均有了明显提高。基于 TF-IDF 方法抽取效果相比其他 2 种方法最差,其原因是未考虑非事件句造成效率下降,基于新闻标题(Title)的关键事件主题句抽取方法效果好于 TF-IDF 方法,除考虑句子本

身特征外,还考虑句子与标题之间主题相关性和语义相关性。本文方法的抽取效果比 Title 方法仍有提高,主要原因是综合考虑事件之间的相互关联关系,加入了影响句子权重的 3 个影响因子,同时对事件句进行真伪辨别。

5 结束语

本文针对原子事件级别的事件抽取粒度过细、主题级别的事件抽取粒度较粗、精确提取事件信息的效率较差的问题,提出一种新的新闻事件主题句提取方法。该方法利用 ICTCLAS 中文处理模块对新闻文本进行处理,如分句、停用词过滤、分词等;采用词语位置加权的 TextRank 方法提取关键词,并计算文本句子关键词的 MI 值,将句子划分为事件句和非事件句;利用 TextRank 模型对事件句建立关系图模型,引入位置关系、句子相似度和关键词覆盖率 3 个影响因子,计算事件句之间的影响力权重,从而迭代算出权重最高的句子作为关键事件主题句。实验结果表明,针对单文档的关键事件主题句提取,与基于 TF-IDF 和 Title 的关键事件主题句抽取方法相比,本文方法的抽取效果更好。

参考文献

[1] 韩客松,王永成,沈 洲,等. 三个层面的中文文本主题自动提取研究[J]. 中文信息学报,2001,15(4):20-26.

[2] 马颖华,王永成,苏贵洋. 一种基于字共现频率的汉语文本主题抽取方法[J]. 计算机研究与发展,2003,40(2):874-878.

[3] 温有奎,温 浩,徐瑞颐,等. 基于创新点的知识元挖掘[J]. 情报学报,2005,24(6):664-668.

[4] 张云涛,龚 玲,王永成. 基于综合方法的文本主题句的自动抽取[J]. 上海交通大学学报,2006,40(5):771-774,782.

[5] Biadsy F, Hirschberg J, Filatova E. An Unsupervised Approach to Biography Production Using Wikipedia[C]// Proceedings of Meeting of the Association for Computational Linguistics. Columbus, USA: [s. n.], 2008:807-815.

[6] 何 维,王 宇. 基于句子关系图的网页文本主题句抽取[J]. 现代图书情报技术,2009(3):57-61.

[7] 刘 茵,李弼程,郭映月. 一种基于聚类算法的主旨句提取方法[J]. 情报学报,2008,27(1):49-55.

[8] 张其文,李 明. 文本主题的自动提取方法研究与实现[J]. 计算机工程与设计,2006,27(15):2744-2746,2766.

[9] 薛扣英,原 盛,张心严. 基于 WFC 和 MI 的主题句提取方法[J]. 计算机工程,2009,35(20):184-186.

[10] 王 伟,赵东岩,赵 伟. 中文新闻关键事件的主题句识别[J]. 北京大学学报(自然科学版),2011,47(5):789-796.

[11] 孔 胜,王 宇. 基于句子相似度的文本主题句提取算法研究[J]. 情报学报,2011,30(6):605-609.

[12] 葛 斌,李芳芳,李 阜,等. 基于无向图构建策略的主题句抽取[J]. 计算机科学,2011,38(5):181-185.

[13] 顾益军,夏 天. 融合 LDA 与 TextRank 的关键词提取研究[J]. 现代图书情报技术,2014,30(7):41-47.

[14] 王 力,李培峰,朱巧明. 一种基于 LDA 模型的主题句抽取方法[J]. 计算机工程与应用,2013,49(2):160-164.

[15] 夏 天. 词语位置加权 TextRank 的关键词抽取研究[J]. 现代图书情报技术,2013,237(9):30-34.

[16] 方 康,韩立新. 基于 HMM 的加权 Textrank 单文档的关键词抽取算法[J]. 信息技术,2015(4):114-116,120.

编辑 顾逸斐

(上接第 218 页)

[8] Castro-Schez J J, Miguel R, Vallejo D, et al. A Highly Adaptive Recommender System Based on Fuzzy Logic for B2C e-commerce Portals[J]. Expert Systems with Applications, 2011, 38(3):2441-2454.

[9] 郝立燕,王 靖. 基于项目流行度的协同过滤 TopN 推荐算法[J]. 计算机工程与设计,2013,34(10):3497-3501.

[10] 李春喜. 一种混合模式电子商务推荐技术的研究[D]. 苏州:苏州大学,2010.

[11] 韩 丽. 社交网络中的信任推荐和好友搜索过滤算法[D]. 秦皇岛:燕山大学,2012.

[12] 高全力,高 岭,杨建锋,等. 融合影响因子的加权协同过滤算法[J]. 计算机工程,2014,40(8):38-42.

[13] 嵇晓声,刘宴兵,罗来明. 协同过滤中基于用户兴趣的相似性度量方法[J]. 计算机应用,2010,30(10):2618-2620.

[14] 严冬梅,鲁城华. 基于用户兴趣度和特征的优化协同过滤推荐[J]. 计算机应用研究,2012,29(2):497-500.

[15] 党 博,姜久雷. 基于评分差异度和用户偏好的协同过滤算法[J]. 计算机应用,2016,36(4):1050-1053.

[16] 朱丽中,徐秀娟,刘 宇. 基于项目和信任的协同过滤推荐算法[J]. 计算机工程,2013,39(1):58-62.

编辑 顾逸斐