【信息科学与控制工程】 doi: 10.11809/scbgxb2017.06.019

基于 LexRank 的中文单文档摘要方法

刘海燕,张 钰

(装甲兵工程学院 信息工程系,北京 100072)

摘要:针对目前中文自动文本摘要方法主要使用基于特征词词频、基于物理位置以及聚类统计的方法准确率较低、不适合单文档摘要,提出了一个改进的中文单文档摘要方法;该方法将 LexRank 算法与 VSM 相结合,充分考虑特征词、特征句、特征段位置等因素;利用 java 语言对其进行实验测试,实验结果表明:改进的自动文本摘要方法和传统摘要方法相比能够更好的实现对文章的自动摘要;该摘要方法可应用到信息挖掘、信息分类、信息索引等领域,在现今信息化的社会,具有较高的现实意义及实用使用价值。

关键词:文本摘要;LexRank 算法;VSM;测评

本文引用格式:刘海燕,张钰. 基于 LexRank 的中文单文档摘要方法[J]. 兵器装备工程学报,2017(6):85 - 89.

Citation format:LIU Hai-yan, ZHANG Yu. Chinese Single Document Summarization Based on LexRank [J]. Journal of Ordnance Equipment Engineering,2017(6):85 - 89.

中图分类号:TP393 **文献标识码**:A **文章编号**:2096 - 2304(2017)06 - 0085 - 05

Chinese Single Document Summarization Based on LexRank

LIU Hai-yan, ZHANG Yu

(1. Department of Information Engineering, Academy of Armored Force Engineering, Beijing 10072, China)

Abstract: Chinese automatic text summarization mainly uses the method based on key words' frequencies, the method based on physical location and the clustering statistics method at present, but the accuracy is low, besides, they are not suitable for the single document summarization. To solve these problems, an improved Chinese single document summarization is mentioned. This improved method combines LexRank algorithm with VSM, and takes full account of factors, such as key words, key sentences and key paragraphs' physical location. Then this designed method is tested by using java language. It turned out that the improved automatic text summarization method can achieve the automatic summarization of the article better than the traditional abstract ones. This summarization method can also be applied in the fields of information mining, information classification, information indexing and else. In this information society, this method has a high practical significance and practical value.

Key words: summary; LexRank algorithm; VSM; evaluation

随着大数据时代的到来,互联网上的数据呈现出爆炸性增长。面对网上纷繁的信息,对于现在的人们来说,能够快速过滤出自己所需要的信息变得格外重要。自动文本摘要能够满足人们这一需求,具有很大的实际应用价值。

自动文本摘要技术在国外的研究起源比较早。在 20 世纪 50 年代,IBM 公司的 H. P. Luhn[1] 开启了研究的先河,他在 1958 年进行了自动摘要系统实验,标志着自动摘要技术的诞生。相比之下,国内自动文本摘要技术的研究起步较晚,1988 年,上海交通大学的王永成[2] 教授研制出 SJTUAA 系统,该系统够较好地实现中文文本自动摘要。近些年来中

收稿日期:2017 - 03 - 05;**修回日期**:2017 - 03 - 30

作者简介:刘海燕(1970—),女,博士,教授,硕士生导师,主要从事信息安全与网络对抗研究。

通讯作者:张钰(1994—),女,硕士研究生,主要从事信息安全与网络对抗研究。

文自动文本摘要技术的研究日益火热。目前,主要使用[3]基于特征词词频、基于物理位置以及聚类统计的方法,这些方法一般不考虑句子之间、段落之间的相似关系,并且主要应用在多文档摘要生成中,不适合单文档摘要领域。

本文提出了一个基于 LexRank 算法,结合 TF-IDF 算法、结合 VSM,并考虑特征词、特征句、特征段位置的适合单文档的中文自动文本摘要系统,能够快速且较准确地生成文本摘要。

1 LexRank 算法

LexRank 算法[4]是密西根大学的 Gunes Erkan 和 Dragomir R Radev 提出的一种基于图论的自然语言处理方法,主要通过句子之间相似度的判断对文本、词汇进行分类。如图 1 所示,用于自动文本摘要时,LexRank 算法对文章中的句子进行处理,将句子作为节点构造出一个标量图,节点间的连线代表两个句子的相似程度。如果两个句子无关,则两个句子所代表的节点间就没有连线;两个句子相似程度越大,节点间的连线就越粗。在对每个句子进行关键句评分时,要充分考虑每个句子所对应节点的连线数量以及连线粗细,即句子的核心性与相关程度大小。最终按照评分,根据一定阈值,选择其中分数较高的句子作为文章的关键句。

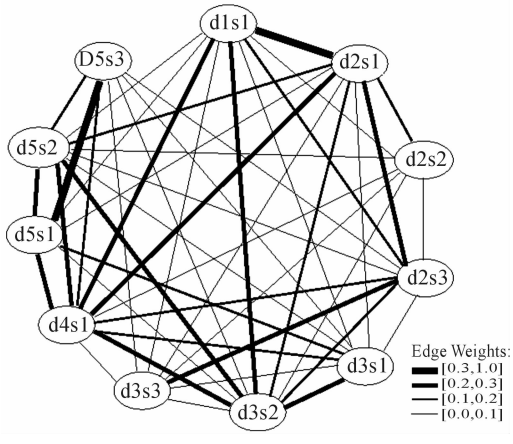

图 1 文本摘要中 LexRank 算示意图

和基于词频的算法相比,LexRank 算法采用基于图的方法,更能有效地考虑句子之间相似度,排除了噪声句对摘要结果的影响。但是单一的 LexRank 算法只是对句子间的相似度进行计算比较,没有考虑在文章中各个自然段落之间的关系。在本文设计的中文单文档摘要系统中,将 LexRank 算法与 VSM 相结合,并将段落之间的关系考虑进去。

2 改进的中文单文档摘要系统

改进后的中文摘要系统流程如图 2 所示,该系统在 LexRank 算法的基础上,充分考虑自然段落相似关系、句子相似关系、句子段落的物理位置等因素,可用于单文档摘要的生成。

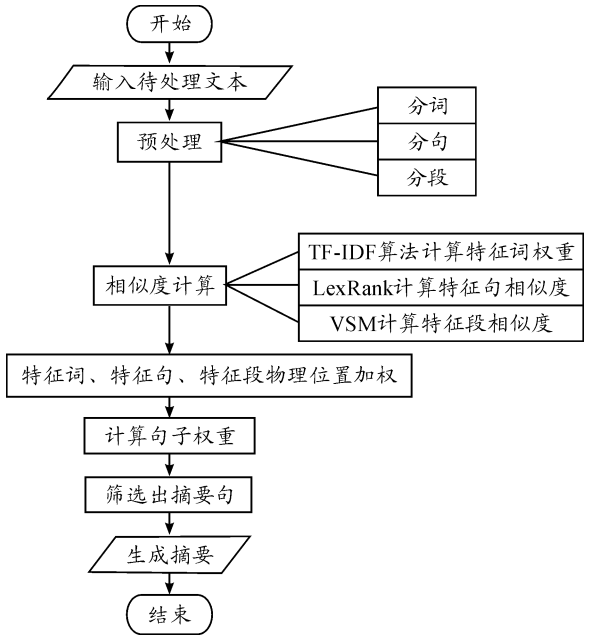

图 2 中文单文档摘要流程

2.1 预处理

面对一篇完整的文档,首先要将其文字转化成可进行数学计算的模型形式,即首先对其进行预处理,把文章进行分词、分句、分段。分句和分段分别根据文章的标点符号以及回车字符就可判断,难点主要在于分词处理。

在现阶段,中文分词技术主要分为 3 种[5]:基于字符串匹配的分词方法、基于理解的分词方法、基于统计的分词方法。本文选取第一种方法。采用大型的语料库对输入的需要测试的文本进行词语比对,然后对其进行分割词汇操作。这种方法对于英文同样适用,只需在语料库中录入英文的语料库即可对英语进行分词。

2.2 TF-IDF 算法计算权重

本文在计算特征词权值时使用了词频-逆文档频率 TF-IDF(Term Frequency-Inverse Document Frequency)[6]算法。其中:TF 为词频,用来计算文档中词语出现的频率;IDF 为逆文档频率,用来排除一些副词、介词等无意义的高频词语。

计算词频 TF 使用的公式如下:

$$TF_i = \frac{n_i}{\sum_k n_k} \tag{1}$$

式中:n_i 为第 i 个词语出现的次数;$\sum_k n_k$ 为文章中所有词语出现的次数。

计算逆文档频率 IDF 所采用的公式如下式所示:

$$IDF(t,D) = \log\left(\frac{N}{n_t}\right) \tag{2}$$

其中,t 表示被测试的词语,D 表示总文档集,N 表示文档的

总个数,n_t 表示含有被测词 t 的文档数量。在单文档摘要系统中,N 则代表句子的总个数,n_t 代表含有被测词 t 的句子数量。n_t 越大,表示被测词 t 的新颖度越低,多为无意义的虚词。式(2)可以看出 n_t 越大,IDF 值越小,因此可以体现词语具有实际意义程度。

为了达到综合考虑的效果,将 TF 与 IDF 二者评分相乘,即最后的单个词语权值如下:

$$W = TF \cdot IDF \tag{3}$$

和传统计算词频求特征值相比,采用 TF-IDF 算法能够有效排除虚词等无实义词的干扰,提高权值计算的准确程度。

2.3 VSM 计算段落相似度

向量空间模型 VSM(Vector Space Model)[7] 是常用的相似度计算模型,在自然语言处理中有着广泛的应用,通常应用在多文档摘要中。如图 3 所示,两个文本向量的夹角表示的就是它们的相似程度,夹角越小证明两篇文档越靠近,即相似度越大[8]。

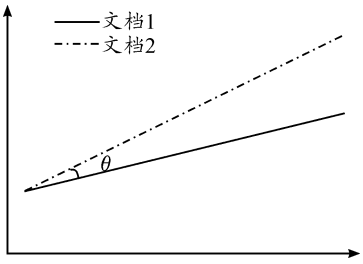

图 3 VSM 文档相似度比较

多文档摘要算法中,VSM 计算公式如下:

$$Sim(T_1,T_2) = \cos(\theta) = \frac{\sum_{k=1}^{N}(w_{1k} \cdot w_{2k})}{\sqrt{(\sum_{k=1}^{N} w_{1k}^2)(\sum_{k=1}^{N} w_{2k}^2)}} \tag{4}$$

式中:T_1 为文档 1;T_2 为文档 2;w_{1k} 为文档 1 中的第 k 个特征词的权重;w_{2k} 为文档 2 中的第 k 个特征词的权重。

在本文设计的自动文本摘要系统中,利用 VSM,将各个段落视为小文档,利用式(4)进行段落相似度计算,如下所示:

$$Sim(P_1,P_2) = \frac{\sum_{k=1}^{N}(W_{1k} \cdot W_{2k})}{\sqrt{(\sum_{k=1}^{N} W_{1k}^2)(\sum_{k=1}^{N} W_{2k}^2)}} \tag{5}$$

式中:P_1 为段落 1;P_2 为段落 2;W_{1k} 为段落 1 中的第 k 个特征词的权重;W_{2k} 为段落 2 中的第 k 个特征词的权重。

由于向量空间的多维性,可以将特征句、特征句权值和特征句所在段落的权值以向量形式表现。将特征句权值、段落权值初始值均记为 0,通过循环迭代计算,段落权值累加直至运算结束,保存在向量中留作评判句子最终权重的一个因素。

和其他算法相比,VSM 具有将特征词、特征句、特征段以及权值构建成对应关系模型的特点,方便对其进行随后的摘要句判别筛选。

2.4 LexRank 算法综合评分

因摘要最后以句子形式组合而成,这里采用 LexRank 算法对文中句子进行相似度计算,本摘要系统设计主要包括 3 步:全文句子的相似度计算、相邻句的相似度计算以及句子最终评分。

1) 全文句子的相似度计算。将作相似度计算的两个句子 S_1 和 S_2 中的词语提取出来,分别记作 $t_{1,1},t_{1,2},\cdots,t_{1,i}$ 和 $t_{2,1},t_{2,2},\cdots,t_{2,j}$,将它们两两作相似度比较,记作 $sim(t_{1,i},t_{2,j})$,将其权值分别记 $w_{1,1},w_{1,2},\cdots,w_{1,n}$ 和 $w_{2,1},w_{2,2},\cdots,w_{2,m}$,这里使用的权值即为前面 TF-IDF 算法求出的特征词权值。所以句子间语义的相似度的如下式所示:

$$SenSim(S_1,S_2) = \frac{\sum_{i=1}^{n} w_{1,i} \cdot m_{1,i}}{\sum_{i=1}^{n} w_{1,i}} + \frac{\sum_{i=1}^{n} w_{2,i} \cdot m_{2,i}}{\sum_{i=1}^{n} w_{2,i}} \tag{6}$$

式中:$m_{1,i}$ 为 $sim(t_{1,i},t_{2,j})$ 中的最大值;$m_{2,i}$ 为 $sim(t_{2,j},t_{1,i})$ 中的最大值。

2) 相邻句相似度计算。在某些情况下,几个不需要的句子互相相关提高其权重,从而对摘要的品质产生负面影响。然而对于核心句子而言,其附近的句子会围绕这个核心展开,即与之相关程度、自身权值均保持较高水平。因此考虑设计计算句子 S 核心程度的公式如下:

$$score(S) = \sum_{S' \in Neighbor(S)} \frac{score(S')}{degree(S')} \tag{7}$$

式中,$score(S)$ 表示句子 S 的核心程度,S' 表示 S 附近的句子,$degree(S')$ 表示 S' 的数量。

3) 句子最终评分。综合句子所在段落权值、句子核心程度以及句子所在物理位置、提示性短语影响等多方面因素,设计句子最终评分为:

$$weight(S) = \alpha \cdot ParaScore + \beta \cdot Score + \chi \cdot OtherScore \tag{8}$$

式中:$ParaScore$ 表示段落权值,$OtherScore$ 表示位置、提示性短语其他因素影响评分,通过分别计算其在被测文档的平均分,再结合实际情况进行加权计算求得。

2.5 摘要句筛选

本文使用统计分析的方法确定摘要筛选的阈值,选出得分最高的句子 S 后,其他句子和 S 的相似度大于阈值则会被视为冗余句筛除。本文定义提取率如下:

$$\text{提取率} = \text{生成摘要字数} / \text{原文字数} \tag{9}$$

由于阈值的范围为 0~1,以 0.1 分度对哈尔滨工业大学的《哈工大信息检索研究室单文档自动文摘语料库》中文档测试,确定阈值范围在 0~0.3。再对 0~0.35 区间以 0.02 分度进行精确阈值测试,结果如图 4 所示。

为了保证摘要的提取率,还要确保摘要语义完整的最大

化,本文根据图4确定选择阈值为0.1。使与 S 相似度大于0.1的被筛除,其他摘要句按照原文顺序排列输出。

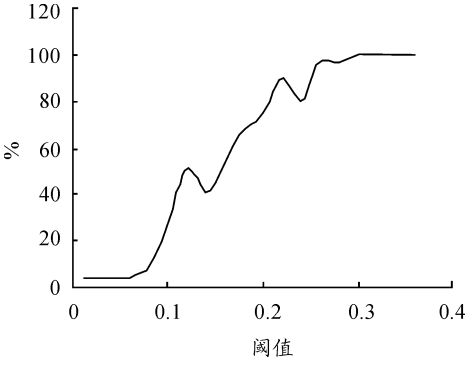

图4 提取率随阈值变化示意图

3 实验与评测

为了验证设计的中文单文档摘要方法的有效性,本文对TF-IDF计算结果、中文摘要结果进行测试,并将结果与原有方法进行了比较。

3.1 TF-IDF 计算结果

本文对设计中各个词语的TF-IDF进行评分检查,观察是否能够实现对文章中各个词语的权值估计。本文对摘自"新浪网"1篇534字的文章进行摘要提取,得到各个词语的TF-IDF分值,并对词语TF-IDF值进行统计,挑选出TF-IDF值大于0的词语并按照其对应的TF-IDF值排序如图5所示。

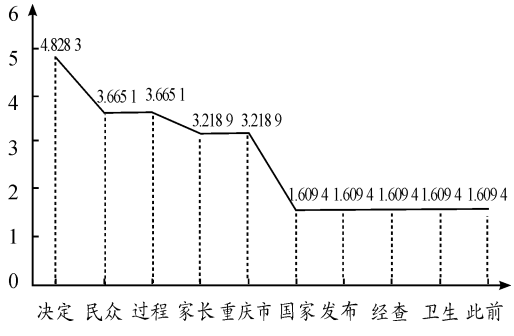

图5 TF-IDF 值折线

为验证此TF-IDF值是否对文本摘要结果产生影响,选择TF-IDF值排名最高的5个词"决定""民众""过程""家长""重庆市"进行研究,观察经过本文设计的中文单文档摘要提取系统后,生成的摘要句中是否包含这几个词。

图6中划线的词就是TF-IDF值最高的词,可以看出,摘要的每个句子都至少包含了一个TF-IDF值前5的词语。因此可见TF-IDF在这个摘要系统中起着重要作用。

3.2 中文摘要结果

本文对哈尔滨工业大学的《哈工大信息检索研究室单文档自动文摘语料库》进行测试,将系统生成的摘要与专家摘要用Edmundson方法[9]加以测评。Edmundson方法比较的是句子,计算公式如下:

$$\text{重合率 } p = \text{匹配句子数} / \text{专家摘要句子数} \times 100\% \tag{10}$$

其中,匹配句子数指的是生成摘要与专家摘要相同的句子的数量。

重庆疫苗事件,用耐心回应纾解焦虑
15日凌晨,重庆市相关调查组发布通报称,由家长见证封存的疫苗经查证,来源渠道规范,运输、储存均符合国家相关规范。此前,5月13日有家长投诉,怀疑重庆市第六人民医院花园路街道社区卫生服务中心注射的自费疫苗被"调包"。
因此,该服务中心是否调包,仍需调查。这也提醒职能部门,回应舆论质疑不是完成式,而应该是一个动态的过程。也许经过调查,疫苗调包不属实,而当地民众或许又有新的疑问,这也需要相关部门直面,而不能选择无视。许多时候,正是因为民众的疑惑得不到解答,合理诉求得不到解决,再加上信息不对称,才愤而走上不合理不合法的维权之路。
"话语权决定主导权,时效性决定有效性,透明度决定公信度。"每逢公共事件,坊间各种问号扎堆。职能部门将问号拉直的过程,就是一个彻查事件的过程,也是一个坦诚回应的过程,更是一个取信于民和提升自身公信力的过程。于此而言,重庆疫苗事件的调查,现在还不能画上句号。

图6 摘要结果

表1 中文单文档摘要提取结果

测试语料	提取率/%	原文的10%专家摘要重合率/%	原文的20%专家摘要重合率/%
奥运	22.63	32.67	28.56
记叙文	13.61	28.57	26.33
说明文	11.47	37.64	29.49
议论文	21.20	39.42	40.33
应用文	9.375	31.76	20.44

从表1以及图7可以看出,在提取率在10%~20%时,本文设计的中文单文档摘要系统对于各种文体均能够有较好的摘要效果,且和原文的10%专家摘要进行比对的效果要好于原文的20%专家摘要,因此本系统对提取最大核心句效果较好。

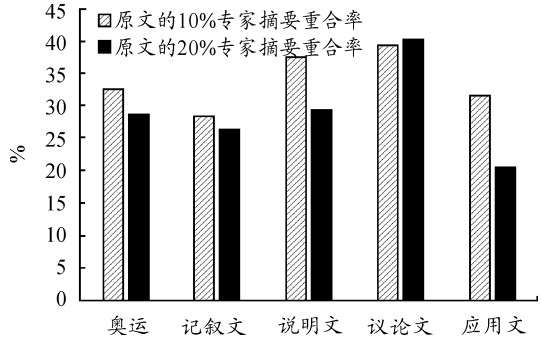

图7 中文摘要提取结果

系统速度方面,为了测试系统速度,选择《百年孤独》的前 4 个章节。由于篇幅过大,使用 txt 文件进行比较分析。进行摘要计算的文件为 91 653 B,摘要结果为 27 247 B,因此提取率为 29.73%,可见该方法能够实现长文本中文单文档提取摘要。而且实验用时小于 15 s,因此证明系统运行比较流畅、高效、快速。

3.3 与原有方法比较

在保证相同提取率的前提下,本文将改进的算法与只使用词频、TF-IDF 算法对《哈工大信息检索研究室单文档自动文摘语料库》中语料进行摘要提取的比较测试,Edmundson 测评结果如图 8、图 9 所示。

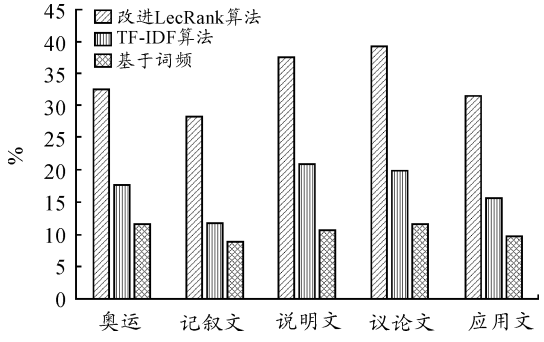

图 8 原文的 10% 专家摘要重合率结果

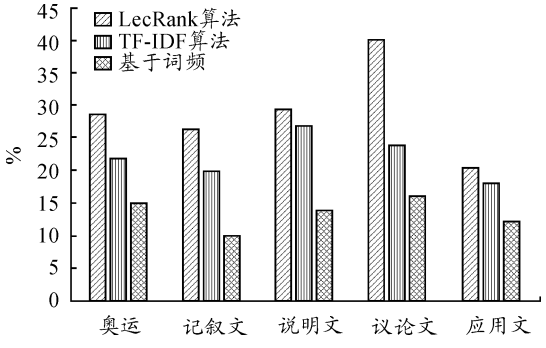

图 9 原文的 20% 专家摘要重合率结果

从图 8、图 9 可以看出,本文设计的改进的基于 LexRank 算法中文单文档摘要系统在各种测试文体中表现均显著优于基于词频、基于 TF-IDF 算法。

4 结论

针对目前中文自动文本摘要提取方法准确度不够高、计算方法速度较慢的问题,本文提出设计一个改进的中文单文档摘要系统。该系统基于 LexRank 算法,将 VSM、TF-IDF 算法结合进去,达到了较好的摘要提取效果。

从实验结果看,它计算速度快,摘要效果良好,和基于词频、TF-IDF 算法相比能够显著提高摘要水平,达到预期的实验设计目的。

本系统的核心思想所涉及的信息挖掘技术、信息分类技术、信息索引技术等在现今信息化的社会,还具有极高的现实意义及实用价值。

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.

[2] WANG Yongcheng. Automatic Extraction of Words from Chinese Textual Data[J]. Journal of Computer Science and Technology, 1987, 2(4):287 – 291.

[3] 胡侠. 自动文本摘要技术综述[J]. 情报杂志, 2010, 29 (8):144 – 147.

[4] GUNES E, RADEV D R. LexRank: Graph-Based Centrality as Salience in Text Summarization[J]. Journal of Artificial Intelligence Research, 2004, 22(10):51 – 54.

[5] 杨阳. 基于 Web 知识的中文分词结果优化[J]. 计算机应用与软件, 2015, 32(12):55 – 58.

[6] AKIKO A. An Information-Theoretic Perspective of TF-IDF Measures[J]. Information Processing and Management, 2002 (7):52 – 57.

[7] 陈炎龙. 基于向量空间模型的英文文本难度判定[J]. 电脑知识与技术, 2010, 12(6):101 – 107.

[8] 刘晓丽. 文本分类检索技术在工程中的应用[J]. 无线电工程, 2008, 38(10):44 – 49.

[9] EDMUNDSON H P. New Methods in Automatic Extracting [J]. Journal of the ACM, 1969, 16(2):264.

[10] 刘星含. 基于互信息的文本自动摘要[J]. 合肥工业大学学报, 2014, 37(10):1198 – 1203.

[11] 曾哲军. 基于连续 LexRank 的多文本自动摘要优化算法研究[J]. 计算机应用与软件, 2013, 30(10):209 – 212.

[12] 纪文倩. 一种基于 LexRank 算法的改进的自动文摘系统[J]. 计算机科学, 2010, 37(5):151 – 154.

(责任编辑 杨继森)