一种改进的 TextRank 关键词提取算法

李志强,潘苏含,戴 娟,胡佳佳
(扬州大学 信息工程学院,江苏 扬州 225000)

摘 要:关键词提取在自然语言处理领域有着广泛的应用,如何准确、快速地从文本中获取关键词信息已经成为文本处理的关键性问题。现有的关键词提取方法很多,但是这些关键词提取方法的准确率和通用性有待提高。因此,提出了一种改进的 TextRank 关键词提取方法,该方法使用 TF-IDF 方法与平均信息熵方法计算文本中词语的重要性,然后根据计算结果得到词语的综合权重。利用词语的综合权重改进 TextRank 算法的节点初始值以及节点概率转移矩阵,通过迭代的方式计算各个节点的权重,直至收敛,从而得到词语的权重信息,选择 top N 个词语作为关键词输出,实现关键词的提取功能。实验结果表明,相较于传统的 TF-IDF 方法和 TextRank 方法,提出的改进后的 TextRank 关键词提取方法有更好的通用性,提取的关键词的准确率更高。

关键词:关键词提取;TF-IDF 算法;TextRank 算法;平均信息熵;自然语言处理
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2020)03-0077-05
doi:10.3969/j.issn.1673-629X.2020.03.015

An Improved TextRank Keyword Extraction Algorithm

LI Zhi-qiang, PAN Su-han, DAI Juan, HU Jia-jia
(School of Information Engineering, Yangzhou University, Yangzhou 225000, China)

Abstract: Keyword extraction is widely used in the field of natural language processing. How to quickly and accurately extract keywords has become the key issue in text processing. At present, there are many methods for keyword extraction, but the accuracy and versatility of them need to be improved. Thus, we propose an improved TextRank keyword extraction method which uses the TF-IDF method and the average information entropy method to calculate the importance of words in the text, and then calculates the comprehensive weight of words based on the calculation results. The initial node weight of the TextRank algorithm and the node probability transfer matrix are improved by using the comprehensive weight of words, and the weights of each node are iteratively calculated until convergence. The weights of the nodes are sorted to obtain the weight information of the words. Then, the Top N words are selected as the keywords. The experiment shows that compared with the traditional TF-IDF method and TextRank method, the improved TextRank keyword extraction method proposed is more general and accurate in keywords extraction.

Key words: Keyword extraction; TF-IDF algorithm; TextRank algorithm; average information entropy; natural language processing

0 引 言

随着信息化的快速普及,网页中的信息数据日益增多。面对海量的文本信息,如何准确、高效地对文章内容进行检索,成为目前的研究热点。对于文本的分析,一般会先从关键词入手,一篇文章的关键词不但可以概括文章的主题,还能反映整篇文章所表达的主要内容与情感倾向。因此,高效、准确地获取关键词,对于文本分类、自动摘要和文本检索至关重要。

近年来,国内外研究人员在关键词提取技术领域展开了大量的研究工作,同时也提出了很多关键词提取算法。其中主要算法有基于隐含主题模型的关键词抽取(LDA)[1]、基于 TF-IDF 词频统计的关键词抽取[2]和基于词图模型的关键词抽取(TextRank)[3]。上述三种算法因其简单易行而应用广泛。

其中,基于隐含主题模型的关键词抽取算法是根据文档和单词的主题分布相似度来计算单词的重要性,由于该方法一般依赖对语料库内容进行训练得到所需信息,因此,这种方法获取到的关键词的质量与训

收稿日期:2019-01-02 **修回日期**:2019-05-09 **网络出版时间**:2019-12-05
基金项目:国家自然科学基金(61070240)
作者简介:李志强(1975-),男,博士,教授,研究方向为 Web 应用、数据分析、量子信息;潘苏含(1994-),男,硕士研究生,研究方向为 Web 数据挖掘、自然语言处理。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191205.1104.002.html

练语料库主题分布密切相关。基于 TF-IDF 算法的词频统计关键词提取(term frequency-inverse document frequency)主要通过对词语出现的频次来判断词语对文章的重要性,是一种成熟的基于统计的提取方法。该方法在关键词提取的过程中过度依赖词频特征,忽略了语义、上下文环境等其他特征。为此,往往会引入其他特征因子进行计算以此减少对词频的依赖。TextRank 算法是基于图的排序算法,利用共现窗口实现部分词语之间的关系构建,对后续关键词进行排序,直接从文本本身中提取关键词。但是该方法没有分析词语重要性的不同是否会影响相邻节点权值转移的问题,并且没有利用文档语料库的整体信息,词语的权重信息并没有实际意义,不能区分连接上的强弱。

为了进一步提高关键词提取的效果与质量,许多学者对上述算法进行改进。郎冬冬等人[4]提出一种基于 LDA 和 TextRank 的文本关键短语抽取方法。顾益军等人[5]结合 LDA 与 TextRank,使候选节点词语的重要性按文档集主题分布进行了非均匀转移。但是结果受训练语料库主题分布影响较大。张瑾等人[6]、谢晋等人[7]考虑了结合词语位置和词跨度的方法对 TF-IDF 权重进行改进,或者利用语义的连贯性,结合词频和词语位置特征进行加权分析[8]。另外也有部分学者引入信息熵[9]的方法。但这些方法在应用过程中都存在一定的问题,比如计算复杂度较高,或者在文章类型和语料规模上需要相当的规模。还有一些学者结合文章综合信息和引用新闻类别因子的方法[10-12],结合了其他特征信息进行加权,在一定程度上能够解决词频依赖问题。但这些方法未考虑关键词的词性以及关键词覆盖度不同所带来的影响。Biswas S K 等人[13]、Yan Ying 等人[14]提出了基于图的关键词提取方法,方法中分别考虑了词语的上下文环境,词语的位置,词语的中心性,词性等特征,修改词语的初始权重等,得到了不错的提取效果。

针对上述问题,基于文献[13-14]的启发,文中综合考虑词语对于单个文档与文档集的重要性来进行关键词抽取。选用 TF-IDF 与平均信息熵综合计算词语对于单文档与文档集的重要性,然后计算出词语的综合权重来改进 TextRank 词汇节点的初始权重以及概率转移矩阵。通过多组实验对比分析,经过改进后的方法能够更高效、准确地获取关键词信息。

1 相关技术

对在关键词提取过程中需要用到的相关算法进行介绍,主要有 TF-IDF 算法与平均信息熵算法。这两种算法主要用于词语对单个文档和文档集的重要性计算。

1.1 TF-IDF 算法

TF-IDF 算法的基本思想是:利用词语频次(term frequency, TF)和逆文档频率(inverse document frequency, IDF)相乘得到词语的权重值[15]。根据 TF-IDF 算法,词语权重 $W_{\text{TF-IDF}}(i)$ 的计算公式如下:

$$W_{\text{TF-IDF}}(i) = \text{TF}_i * \text{IDF}_i \tag{1}$$

$$\text{IDF}_i = \log(N/\text{DF}_i) \tag{2}$$

其中, TF_i 表示词语 i 在文档内容中出现的次数/文档内容的总词数,即词语 i 在文档中出现的频率, N 表示语料库中文档总数, DF_i 表示包含词语 i 的文档数, IDF_i 表示词语 i 的主题表现能力。

根据上述公式可知,如果词语在某一篇文档中出现频率较高,但是在语料库中包含该词的文档数较低,则该词根据 TF-IDF 算法得到的权重值 $W_{\text{TF-IDF}}(i)$ 就越高,也就是说该词语可以在一定程度上表示文章的主题内容。反之,则说明该词语不是重要词语,不能表现文章的主要内容。

1.2 平均信息熵算法

平均信息熵的基本思想是:根据词频在不同文档中出现的频数,结合整体语料库计算所有词语对于单个文档和文档集的重要性,通过平均信息熵可以衡量词语在整个文档集中分布的均衡度。根据平均信息熵算法,词语权重 $W_{\text{Entropy}}(i)$ 的计算公式如下:

$$W_{\text{Entropy}}(i) = 1 - \frac{1}{\log N}\sum_{k=1}^{N}\left(\frac{f_{wk}}{n_w}\log\frac{n_w}{f_{wk}}\right) \tag{3}$$

其中, f_{wk} 表示词 w 在文档 k 中出现的频次, n_w 表示词 w 在整个文档集中出现的频次, N 表示语料库中文档的总数。

如果词 i 在各类别文档中出现频率相当,则其 $W_{\text{Entropy}}(i)$ 的值接近于最小值 0,表示其并不能很好地表示文档的主题内容。反之,如果词语 i 在各类文档中出现频率差别很大,其 $W_{\text{Entropy}}(i)$ 的值接近于最大值 1,表示其对文档主题有很好的表现力。

2 关键词提取算法

2.1 算法描述

传统的 TextRank 算法是将一篇文档转换成一张有向带权的词图模型,是将文本进行分割,分割成基本单元,即词语,每个基本单元看作是一个节点,每个节点之间的边由词节点之间的共现关系决定,而节点的重要性又由相邻节点指向数量决定。TextRank 算法的计算方式如下所示。

$$\text{WS}(V_i) = (1-d) + d * \sum_{V_j \in \text{In}(V_i)} \frac{W_{ji}}{\sum_{V_k \in \text{Out}(V_j)} W_{jk}} \text{WS}(V_j) \tag{4}$$

构建 TextRank 关键词图 $G=(V,E)$,其中 V 为节点集合,E 为节点之间的边集合;$\mathrm{In}(V_i)$ 是节点 V_i 的入度点的集合,即指向节点 V_i 的节点集合;$\mathrm{Out}(V_j)$ 是节点 V_j 的出度点集合,即节点 V_j 指向的所有节点的集合;W_{ji} 是节点 V_j 与节点 V_i 之间边的权重;d 是阻尼系数,一般取值为 0.85,其作用是表示当前节点向其他任意节点跳转的概率,同时能够保证让权重能够稳定的传递至收敛,最终计算每个词语的权重并进行排序,选取 top$N$ 作为 N 个关键词并输出。

传统 TextRank 算法中,每个节点初始权重为 1 或 $1/N$,即节点的初始权重相同,且节点权值均匀转移。但是通过研究发现,基于词语重要性加权词语转移概率的方法可以有效改进关键词提取的效果[16-17]。因此,文中选用 TF-IDF 和平均信息熵两个特征来计算词语的权重,用计算得到的综合特征信息来改进 TextRank 词汇节点的初始权重大小以及概率转移矩阵。

选取任意一个词 i,定义词 i 的综合权重计算方式如下:

$$W_{\mathrm{Weight}}(i)=\frac{1}{2}W_{\mathrm{TF-IDF}}(i)+\frac{1}{2}W_{\mathrm{Entropy}}(i) \tag{5}$$

其中,$W_{\mathrm{TF-IDF}}(i)$ 是词语通过 TF-IDF 计算得到的权重值,$W_{\mathrm{Entropy}}(i)$ 是词语的平均信息熵权值。

根据 TextRank 算法,节点间的转移概率计算公式为:

$$W(V_j,V_i)=\frac{W_{\mathrm{Weight}}(V_i)}{\sum\limits_{V_k\in \mathrm{Out}(V_j)}W_{\mathrm{Weight}}(V_k)} \tag{6}$$

根据式(6)可知,节点的权重迭代公式如下:

$$\mathrm{WS}(V_i)=(1-d)+d*\sum_{V_j\in \mathrm{In}(V_i)}W(V_j,V_i)\mathrm{WS}(V_j) \tag{7}$$

其中,$W(V_j,V_i)$ 表示节点 V_j 到节点 V_i 边的转移概率,即通过式(6)计算所得,$W_{\mathrm{Weight}}(V_i)$ 表示由式(5)得到的综合权重值。

2.2　关键词提取

基于改进的 TextRank 关键词提取算法的提取流程如图 1 所示。

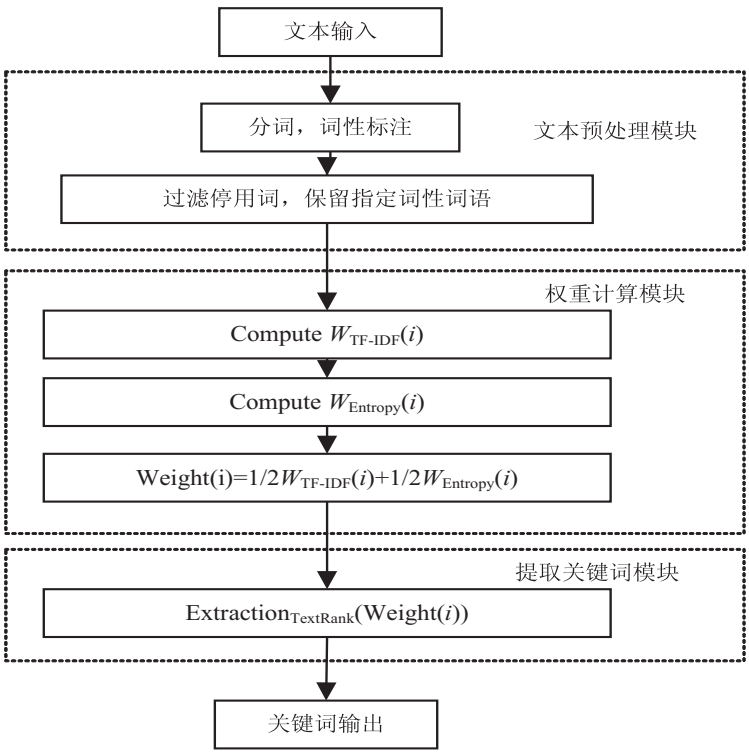

图 1　基于改进的 TextRank 关键词提取算法的提取流程

输入需要提取关键词信息的文本内容,关键词提取步骤如下:

第一步:文本预处理。对文本中的内容进行分词,词性标注,只保留名词、专有名词、动词、形容词和副词,并删除文本中的停用词。

第二步:权重计算。计算得到文本中每个词语的 $W_{\mathrm{TF-IDF}}$,W_{Entropy},并计算综合权重 W_{Weight}。

第三步:关键词提取。构建基于词语综合权重的加权节点初始值及节点概率转移矩阵改进的 TextRank 模型,最终计算选择前 N 个权重比较大的词语作为关键词并输出。

3　实验结果及分析

文中的实验数据是在各大门户网站中随机抽取的

包含多种主题的新闻数据、新闻内容字数、主题等信息。将每一篇新闻保存为一个文档,共 500 个文档组成语料库。对于数据集,采用多人人工交叉标注的形式提取关键词,每一个文档分别提取 5,7,10 个关键词。

对于相同的数据集,提取不同数量的关键词,实验中分别采用传统的 TF-IDF 算法、TextRank 算法和改进的 TextRank 算法(共现窗口大小一个为 5,一个为 7)进行交叉对比。采用精度 P(Precision)、召回率 R(Recall)和 F_1 值(F_1-Measure)作为评价关键词提取的性能指标,指标计算公式如下:

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap T_i|}{|P_i|} \tag{8}$$

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap T_i|}{|T_i|} \tag{9}$$

$$F_1 - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

3.1 实验结果

实验结果如表 1 所示。根据表 1 的实验结果分析可以发现,文中提出的方法在关键词提取方面优于传统的 TF-IDF 和 TextRank 算法。

表 1 实验结果对比 %

方法	关键词个数	Precision	Recall	F_1-Measure
TF-IDF		50.42	24.73	33.18
TextRank-span=5		52.71	26.48	35.25
TextRank-span=7	5	53.54	26.88	35.79
文中算法-span=5		55.23	30.83	39.57
文中算法-span=7		55.45	31.32	40.03
TF-IDF		42.08	29.85	34.92
TextRank-span=5		48.74	34.21	40.2
TextRank-span=7	7	49.91	34.17	40.57
文中算法-span=5		56.21	32.16	40.91
文中算法-span=7		56.55	33.18	41.82
TF-IDF		35.44	34.72	35.08
TextRank-span=5		41.43	40.87	41.15
TextRank-span=7	10	42.37	40.22	41.27
文中算法-span=5		55.93	34.26	42.49
文中算法-span=7		57.12	35.66	43.91

3.2 实验分析

实验结果交叉对比如图 2 所示。

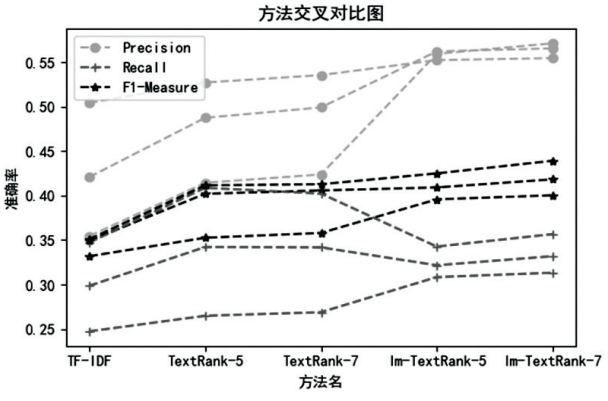

图 2 实验结果交叉对比

观察对比图可以发现,随着提取的关键词数量的增加,传统的 TF-IDF 算法与 TextRank 算法的准确率呈现下降的趋势;而文中提出的改进的 TextRank 算法,随着关键词提取的数量变化,准确率相对稳定。因为根据词语的综合权重进行计算,关键词相对较为集中突出,权重相对较大。

此外,可以发现,传统的 TextRank 算法的共现窗口大小对准确率也有影响。因为共现窗口的大小决定可权重转移概率矩阵的稠密,从而影响关键词的提取结果。当共现窗口为 7 的时候,提取的关键词准确度要比共现窗口为 5 时高一些。

总之,文中提出的改进方法,在精度 P(Precision)和 F_1 值(F_1-Measure)方面均高于传统方法,召回率(Recall)基本相当。这个基本可以证明,该方法在关键词信息获取方面较传统的 TF-IDF 和 TextRank 算法更加高效、准确。

4 结束语

一篇文章的关键词不但可以概括文章的主题,还能反映整篇文章所表达的主要内容与情感倾向,所以对于提取的关键词信息的准确率有较高的要求,这样

才能够相对准确地表达文本的主题内容。因此,高效、准确地提取关键词信息,对于文本分类、自动摘要和文本检索至关重要。

提出了一种基于 TextRank 的改进算法。针对传统 TextRank 算法没有考虑词语本身的重要性,以及文档整体信息等不足之处,该算法选用 TF-IDF 算法与平均信息熵算法计算词语的重要性,通过计算词语的重要性对不同词语赋予不同的权重,根据计算得到的词语权重改进 TextRank 算法词汇节点的初始权重以及概率转移矩阵。该算法提高了关键词提取的准确度,并且操作简单,无须进行训练和人工干预,具有较强的通用性,能够满足对于一般文章的关键词提取需求。同时,可以结合更多的词语特征、词语上下文的语义环境等对该算法进行完善,这也是接下来研究的主要方向。

参考文献:

[1] BLEI D M, NG A Y, JORDAN M. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(4-5): 993-1022.

[2] LI J, FAN Q, ZHANG K. Keyword extraction based on tf/idf for Chinese news document[J]. Wuhan University Journal of Natural Sciences, 2007, 12(5): 917-921.

[3] MIHALCEA R, TARAU P. TextRank: bringing order into texts[C]//Proceedings of the conference on empirical methods in natural language processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004: 404-411.

[4] 郎冬冬,刘晨晨,冯旭鹏,等. 一种基于 LDA 和 TextRank 的文本关键短语抽取方案的设计与实现[J]. 计算机应用与软件,2018,35(3):54-60.

[5] 顾益军,夏　天. 融合 LDA 与 TextRank 的关键词抽取研究[J]. 现代图书情报技术,2014,30(z1):41-47.

[6] 张　瑾. 基于改进 TF-IDF 算法的情报关键词提取方法[J]. 情报杂志,2014,33(4):153-155.

[7] 谢　晋. 基于词跨度的中文文本关键词自动提取方法[J]. 现代物业·现代经济,2012,11(4):108-111.

[8] 胡学钢,李星华,谢　飞,等. 基于词汇链的中文新闻网页关键词抽取方法[J]. 模式识别与人工智能,2010,23(1):45-51.

[9] 李　航,唐超兰,杨　贤,等. 融合多特征的 TextRank 关键词抽取方法[J]. 情报杂志,2017,36(8):183-187.

[10] 刘奇飞,沈炜域. 基于 Word2Vec 和 TextRank 的时政类新闻关键词抽取方法研究[J]. 情报探索,2018(6):22-27.

[11] 王雍凯,毛存礼,余正涛,等. 基于图的新闻事件主题句抽取方法[J]. 南京理工大学学报,2016,40(4):438-443.

[12] YAN Y, HE L, MENG Q. Exploration and improvement in keyword extraction for news based on TFIDF[J]. Energy Procedia, 2011(13): 3551-3556.

[13] BISWAS S K, BORDOLOI M, SHREYA J. A graph based keyword extraction model using collective node weight[J]. Expert Systems with Applications, 2018, 97: 51-59.

[14] YAN Y, TAN Q, XIE Q, et al. A graph-based approach of automatic keyphrase extraction[J]. Procedia Computer Science, 2017, 107: 248-255.

[15] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,29(z1):167-170.

[16] 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(9):30-34.

[17] 周锦章,崔晓晖. 基于词向量与 TextRank 的关键词提取方法[J]. 计算机应用研究,2019,36(4):1051-1054.

(上接第 76 页)

Scientific Research and Management, 2014, 35(2): 95-98.

[13] 王永辉. 基于 Nginx 高性能 Web 服务器性能优化与负载均衡的改进与实现[D]. 成都:电子科技大学,2015.

[14] 林丽丽. 使用高性能 Web 服务器 Nginx 实现开源负载均衡[J]. 大众科技,2010(7):37-38.

[15] BITTAU A, BELAY A, MASHTIZADEH A, et al. Hacking blind[C]//2014 IEEE symposium on security and privacy. San Jose, CA: IEEE, 2014: 227-242.

[16] 马原龙. Nginx 负载均衡技术研究[D]. 重庆:重庆邮电大学,2016.

[17] CHI X, LIU B, NIU Q, et al. Web load balance and cache optimization design based nginx under high-concurrency environment[C]//2012 third international conference on digital manufacturing and automation (ICDMA). [s. l.]: [s. n.], 2012: 62-67.

[18] 李军锋,何明昕. 高并发 Web 航空票务秒杀系统的设计与实现[J]. 计算机工程与设计,2013,34(3):778-782.