融合多因素的 TFIDF 关键词提取算法研究

牛永洁,田成龙
(延安大学 数学与计算机学院,陕西 延安 716000)

摘 要:为了能更加准确、快速地提取文本中的关键词,首先需要对待提取的文本进行数据清洗,去掉其中的噪声数据,接着对文本进行分词操作,在去掉停用词的基础上,综合考虑词语的位置、词性、词语关联性、词长和词跨度等因素,将这些因素与经典的 TFIDF 关键词提取算法相结合,采用不同权重的方法得到最终的词语权重,按照词语权重从大到小取得前 5 个词作为文本的关键词。以本校图书馆提供的 8 045 篇《红色中华》新闻为源数据,从准确度、召回率及 F_1 值三个指标对文中算法、经典的 TFIDF 算法和专家标注进行对比,发现文中算法在三个指标上均优于经典的 TFIDF 算法,与专家标注比较接近。

关键词:TFIDF 算法;词位置;词性;词语关联;词长;词跨度
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2019)07-0080-04
doi:10.3969/j.issn.1673-629X.2019.07.016

Research on TFIDF Keyword Extraction Algorithm Based on Multiple Factors

NIU Yong-jie, TIAN Cheng-long
(School of Mathematics & Computer, Yan'an University, Yan'an 716000, China)

Abstract: In order to extract the key words in the text more accurately and quickly, the first step is to clean the extracted text, remove the noise data, and then perform word segmentation on the text. On the basis of removing the stop words, the word location, part of speech, word relevance, word length and word span are considered comprehensively. These factors are combined with the classic TFIDF key word extraction algorithm. The final word weight is obtained by using the method of different weights, and the first five words are taken as the key words in the text according to the weight of words from large to small. Based on the news of the 8 045 "Red China" provided by the library, by comparing the algorithm proposed, the classical TFIDF algorithm and expert annotation from three indexes of accuracy, recall rand F_1, it is found that the algorithm proposed is superior to the classical TFIDF algorithm in three indexes and is close to expert annotation.

Key words: TFIDF; word position; part of speech; word correlation; word length; word span

0 引 言

随着数据时代的到来,各行各业都积累了大量的数据,人们迫切希望从这些数据中发现有趣的知识。自然语言处理研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理融合了语言学、计算机科学、数学等学科,针对非结构化的文本信息进行处理,其中关键词的提取是自然语言处理中的基础与核心技术,在信息检索、文本分类、文本聚类、信息匹配、话题跟踪、自动摘要、人机对话等领域有广泛的应用[1-4]。

目前针对文本关键词的提取,为了取得良好的效果,大都采用专家标准的方法,但是面对日益增多的海量文本信息和迫切的应用需求,人工标注已经显得力不从心。于是借助计算机自动进行关键词提取的方法受到了越来越多的重视,已经成为自然语言处理领域的一个研究热点[5-7]。

关键词抽取方法按照是否进行监督学习分为监督性和非监督性两大类。通过训练数据构建学习模型,进而判断词语是归属于关键词类别还是非关键词类别,属于典型的有指导学习方法。有指导学习需要事

收稿日期:2018-09-13 **修回日期**:2019-01-04 **网络出版时间**:2019-03-21
基金项目:国家社会科学基金项目(18BTQ042);国家级大学生创新创业训练计划项目(201710719024)
作者简介:牛永洁(1977-),男,硕士,副教授,CCF 高级会员(09256S),研究方向为数据挖掘、大数据;田成龙(1998-),男,专业为计算机科学与技术。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190321.0942.080.html

先标注高质量的训练数据,人工预处理的代价较高。非监督学习因为无需对数据进行训练,实现快捷,仅需要文本自身的信息就能进行等优点被广泛采用,非监督关键词抽取的主流方法可归纳为三种:基于 TFIDF 统计特征的关键词抽取、基于主题模型的关键词抽取和基于词图模型的关键词抽取。这些方法都有自己的优缺点[8]。

文中主要针对 TFIDF 展开研究,综合考虑文本信息中词语的位置、词性、词语关联性、词长和词跨度 5 种影响因素,对每一种影响因素赋予一定的权重,最后加和得到最终的词语权重,获得权重最大的前 5 个词语作为文本的关键词。与经典的 TFIDF 方法及人工标注进行对比,发现文中算法在精确度、召回率和 F_1 值都优于经典的方法,更加接近人工标注,值得推广应用。

1　相关技术

相关技术主要包含 TFIDF、词语的位置、词性、词语关联性、词长和词跨度 6 个方面。设定一个文本集合 D,集合中包含 N 个文本,每个文本 T 包含标题 title 和内容 content 两部分。content 内容由若干段落 segment 组成,段落由换行回车键进行分割。每个段落包含若干句子 sentence,句子由若干词语 word 组成。句子由标点符号“。”、“!”、“?”、“……”进行分割。

1.1　TFIDF 算法

TFIDF 算法处理的对象是文本的 content 部分[9-10],其中每个词语 word 的权重由式 1 进行计算。

$$W_{\mathrm{tf}}(i) = \mathrm{tf}_i * \mathrm{idf}_i \tag{1}$$

其中,$W_{\mathrm{tf}}(i)$ 表示第 i 个词语使用 TFIDF 方法得到的权重;tf_i 表示该词的词频,词频为该词在 content 中出现的次数与 content 中词语总数之比;idf_i 表示逆文档频率,计算方法为:

$$\mathrm{idf}_i = \log\left(\frac{N}{\mathrm{df}_i} + \beta\right) \tag{2}$$

其中,N 为文档总数;df_i 为文档中出现词语 i 的文档数;β 为一个经验常数,一般取 0.01、0.1、1,文中取数值 1。

TFIDF 的计算表明,如果一个词语在文本 content 中出现的次数越多但是在集合 D 中包含该词语的其他文本数量越少,该词语成为文本关键词的权重越大,文中采用 W_tfidf 表示词语的权重。

1.2　词语的位置

根据文献[4,8],文本的标题 title 一般会尽可能包含文本的中心思想,所以出现在标题中的词语成为关键词的概率最大,另外一个文本的第一段往往是全文的初步概括,也能最大限度地体现文章的主旨,所以对出现在第一段中的词语也需要增加权重,末段往往是对全文的总结,因此也需要对出现在末段的词语增加权重。每段内容的首句往往是本段内容的纲领,所以出现在每段第一句中词语的权重也应该适当重视。词语位置的权重设置如表 1 所示。

表 1　词语位置权重设置

出现位置	权重名称	权重设置
标题	W_title	6
首段	W_fseg	5
末段	W_eseg	4
首句	W_fsen	3
其他	W_locother	1

1.3　词　性

汉语词性可以分为实词和虚词。实词包含:名词、动词、形容词、数词、量词和代词。虚词包括:副词、介词、连词、助词、叹词、拟声词。关键词词性分布一般是名词或名词性短语为主,其次是动词,最后是数词、副词和其他修饰词等[11]。考虑词性特征可以有效避免传统采用语言学方法的缺陷[12-15],词性的权重设置如表 2 所示。

表 2　词性权重设置

词性	权重名称	权重设置
名词性	W_posn	6
动词性	W_posv	4
形容词、副词	W_posa	3
数词、量词	W_posm	2
其他	W_posother	1

1.4　词语关联性

汉语语言的词语之间的关联度在全局上显示出高度的连接性,同时在局部具有高度的聚集性。根据自然语言具有的关联特性,可以作为基本特征进行关键词提取。因为在实践中 TFIDF 算法的固有缺陷表现为数据集偏斜,类间、类内分布偏差等。在词语关联度算法方面,由于复杂网络仅仅依靠词语之间的相互关系作为基本特征,忽略了单词的频率特征,容易造成关键词提取的聚集特征不明显,从而引起关键词提取的误差[16-17]。将二者相结合可以互相补充,能够更加全面地描述一个词语的权重。

设 $V = \{v_1, v_2, \cdots, v_n\}$ 为节点集合,(v_i, v_j) 表示节点 $v_i \in V$ 与 $v_j \in V$ 之间的边。设 $G(V,E)$ 是以 V 为节点集合,以 $E \subset \{(v_i, v_j) : v_i, v_j \in V\}$ 为边集合的图,则节点 v_i 的度 D_i 为:

$$D_i = |\{v_i, v_j\} : (v_i, v_j) \in E, v_i, v_j \in V| \tag{3}$$

节点 v_i 的聚集度 K_i 为：

$$K_i = |\{v_j, v_k\} : (v_i, v_j) \in E, (v_i, v_k) \in E, v_i, v_j, v_k \in V| \quad (4)$$

节点 v_i 的聚集系数 C_i 为：

$$C_i = \frac{K_i}{\binom{D_i}{2}} = \frac{2K_i}{D_i(D_i - 1)} \quad (5)$$

对于节点 v_i 计算网络综合特征值 CF_i：

$$CF_i = \frac{\alpha C_i}{\sum_{j=1}^{N} C_j} + \frac{(1-\alpha)D_i}{N} \quad (6)$$

其中，N 表示网络中的节点个数，$0<\alpha<1$，文中取 α 为 0.5。

对于文本中的每一个句子 sentence，将句子 sentence 中的词语作为节点集合，将各个句子所组成的网络连接，合并相同的节点和连边，就形成一个语言网络。根据文献[13]的研究成果，只考虑词关联跨度为 1 和 2，计算每个词语的度 D，聚集度 K 和综合特征值 CF。使用 CF 值作为词语 word 的词关联性权重 W_cf。

1.5 词 长

经过研究发现，一个文本的关键词的词长一般大于 2，所以可以将词长小于 2 的词语过滤掉。关键词词长越长，包含的信息越大，但是关键词词长一般不超过 6，因此也可以将词长大于 6 的词语过滤掉。可以使用式 7 作为词长的权重。

$$W\_len = \frac{词长}{词长 + 4} \quad (7)$$

1.6 词跨度

一个词的跨段落情况说明这个词是描述局部的还是表达全文的。跨段数越多，说明该词越重要，全局性越强。显然，局部关键词不是需要提取的目标，然而在传统 TFIDF 算法中，局部关键词往往会因为其高频优势成为整个文档的关键词，降低了提取关键词的准确率[18]。在提取关键词的过程中，为了体现词语的全局性，利用式 8 来衡量词语的跨度权重。

$$W\_seg = \frac{词语出现的段数}{文本总段数} \quad (8)$$

2 算法步骤

融合多因素的 TFIDF 的算法步骤为：

(1)数据清洗：将文本中的噪声数据清除，比如文本中多余的空格、 、#、*、[、]、【、】等字符。

(2)标记：对文本进行段落识别，标记首段、末段，对文本进行语句识别，标记句子的开始和结束和每段的首句。

(3)分词：对文本进行带有词性的分词，分词结果分为两个集合，分别是标题的分词结果和内容的分词结果。文中采用了北京理工大学海量语言信息处理与云计算工程研究中心的 NLPIR 汉语分词系统进行分词。

(4)停用词过滤：停用词在文本分析中属于一种冗余数据，对文本的主题不具备表达能力，往往具有高频、无意义等特点。例如，“的”、“啊”、“但是”等词语以及标点符号通过去除停用词，能消除对关键词提取的干扰。

(5)词性过滤：将文本中经过分词且词性被标记为介词、连词、助词、叹词、拟声词、语气词等词语过滤掉，这些词通常不可能是关键词，同时会增加后续计算的工作量，所以将这些词过滤掉。

(6)词长过滤：将词长长度小于 2 大于 6 的词语过滤掉。

(7)采用 TFIDF 算法计算每个词语的 W_tfidf。

(8)根据词语的位置计算每个词语的位置权重。

(9)根据词性分别计算每个词的权重。

(10)计算词语的词关联性权重 W_cf。

(11)计算词语的词跨度权重 W_seg。

(12)计算词语的词长权重 W_len。

(13)根据式 9 计算词语的最终权重 W_all。

$$W\_all = (\alpha W\_tfidf + \beta W\_cf + \gamma W\_seg + \delta W\_len) * 位置权重 * 词性权重 \quad (9)$$

其中，α、β、γ、δ 为各种不同权重的加权系数，文中取 α 为 1.5，β 为 1.1，γ 为 0.8，δ 为 0.5。

将计算得到的词语的最终权重按照降序排列，取前 5 个作为一篇文本的关键词。

3 测试及结论

为了衡量关键词提取算法的优劣，往往采用 3 个指标作为衡量的标准，分别是准确率、召回率和 F_1 值，其中准确率和召回率是一对相互矛盾的指标，也就是说准确率如果比较高，但是召回率要低一些，综合这两个指标提出了 F_1 值的概念，如果 F_1 值比较高，则说明算法的效果比较好。

准确率通过式 10 进行计算。

$$准确率 = \frac{Num_{correct}}{Num_{total}} \quad (10)$$

其中，$Num_{correct}$ 表示正确提出的关键词数量；Num_{total} 为总共提出的关键词数量。

召回率通过式 11 进行计算。

$$召回率 = \frac{Num_{correct}}{Num_{actual}} \quad (11)$$

其中，Num_{actual} 为文本实际的关键词数量。

F_1 值综合考虑了准确率和召回率两个指标,通过式 12 进行计算。

$$F_1 = \frac{2 * 准确率 * 召回率}{准确率 + 召回率} \tag{12}$$

本校图书馆对《红色中华》报刊进行了收集和整理,共得到从 1931 年到 1937 年 6 年间的 8 045 篇新闻文章,其中每篇文章都由标题和正文组成,其中部分文章已经通过红色文献研究专家进行了关键词提取和标注工作。8 045 篇文章作为文本的全体样本,每篇文章作为一个文本,按照文中提出的算法进行了关键词提取。通过准确率、召回率和 F_1 值对文中算法、经典的 TFIDF 算法和专家标注进行了对比,结果如表 3 所示。

表 3　算法对比　　%

算法	准确率	召回率	F_1 值
文中算法	80.3	72.4	76.2
经典的算法	57.8	47.2	51.9

通过表 3 可以看出,融合多种因素的文中算法在三个指标上都明显优于经典的 TFIDF 算法,值得推广应用。但是该算法也有不完善的地方,主要表现在计算工作量大,运行时间长,但是如果作为已经整理好的离线数据源,为了提高关键词提取的效果仍然是一种比较好的方法。通过对文中算法和专家标注的结果进行对比,发现该算法仍然有一些缺陷,主要表现为词语组合问题,比如:专家标注的关键词“满洲傀儡政府”,在文中算法中被分为两个词“满洲”和“傀儡政府”,可以看出文中算法的结果一方面受到分词系统的影响,另一方面应该根据词语的关联度进行词语的组合,但是汉语的语法比较灵活,词语组合规则还很难提取和总结,所以词语组合问题还有待于进一步研究。

4　结束语

通过综合考虑词语的位置、词性、词长、词跨度和词语关联度等多种因素对经典的 TFIDF 算法进行了改进,对每个因素的权重进行了加权相加或者相乘的运算,得到一个最终的词语权重,然后取权重值最大的 5 个词语作为文本的关键词,以专家手工标注的关键词为标准,对两种算法进行了对比,发现文中算法效果良好,值得推广应用,同时在研究的过程中也发现了一些不足和缺陷。总而言之,文中算法比较全面地考虑了影响关键词提取的各种因素,具有一定的通用性,能够为其他类似的研究提供思路和参考,具有一定的推广性和借鉴性,同时也为下一步研究指明了方向。

参考文献:

[1] 杨凯艳. 基于改进的 TFIDF 关键词自动提取算法研究[D]. 湘潭:湘潭大学,2015.

[2] 杨　玥. 中文文本主题关键短语提取算法研究[D]. 西安:西安理工大学,2017.

[3] 董　苑. 科技论文查询可视化系统设计与实现[D]. 杭州:浙江工业大学,2017.

[4] HABIBI M, POPESCU-BELIS A. Keyword extraction and clustering for document recommendation in conversations [J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2015, 23(4): 746-759.

[5] 王　洁,王丽清. 多特征关键词提取算法研究[J]. 计算机系统应用,2018,27(7):162-166.

[6] 陈伟鹤,刘　云. 基于词或词组长度和频数的短中文文本关键词提取算法[J]. 计算机科学,2016,43(12):50-57.

[7] ABILHOA W D, DE CASTRO L. A keyword extraction method from twitter messages represented as graphs[J]. Applied Mathematics and Computation, 2014, 240: 308-325.

[8] 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.

[9] 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.

[10] 苏祥坤,吾守尔·斯拉木,买买提依明·哈斯木. 基于词序统计组合的中文文本关键词提取技术[J]. 计算机工程与设计,2015,36(6):1647-1651.

[11] 张红鹰. 中文文本关键词提取算法[J]. 计算机系统应用,2009,18(8):73-76.

[12] 钱爱兵,江　岚. 基于改进 TF-IDF 的中文网页关键词抽取—以新闻网页为例[J]. 情报理论与实践,2008,31(6):945-950.

[13] CHEN Y H, LU J L, MENG F T. Finding keywords in blogs: efficient keyword extraction in blog mining via user behaviors[J]. Expert Systems with Applications, 2014, 41(2): 663-670.

[14] 张建娥. 基于 TFIDF 和词语关联度的中文关键词提取方法[J]. 情报科学,2012,30(10):1542-1544.

[15] PALIWAL R, RANA M S. Content analysis and application of Zipf's law in computer science literature[C]//Proceedings of the 2015 4th international symposium on emerging trends and technologies in libraries and information services. Noida, India: IEEE, 2015: 223-227.

[16] 谢　晋. 基于词跨度的中文文本关键词提取及在文本分类中的应用[D]. 杭州:浙江工业大学,2011.

[17] 罗　燕,赵书良,李晓超,等. 基于词频统计的文本关键词提取方法[J]. 计算机应用,2016,36(3):718-725.

[18] 李静月,李培峰,朱巧明. 一种改进的 TFIDF 网页关键词提取方法[J]. 计算机应用与软件,2011,28(5):25-27.