文章编号：2096-1472(2018)-02-04-03
DOI:10.19644/j.cnki.issn2096-1472.2018.02.002

改进的TFIDF标签提取算法

王 杰[1]，李旭健[2]
(1.山东科技大学，山东 青岛 266590；
2.山东省数字矿山重点实验室，山东 青岛 266590)

摘 要：TFIDF算法作为一种加权算法，在信息检索和数据挖掘等自然语言处理领域发挥了巨大的作用。它的计算模型相对简单，适合大数据并行计算，适用领域广泛，且拥有很好的解释性。基于以上这些特点，本文在TFIDF算法基础之上，利用监督的学习，并通过引入加权因子和词贡献度，来修正TFIDF算法结果权值。利用这个算法可以在自然语言处理中有效地提取特征标签，并且改进后的算法在这一细分领域具有极高准确度。

关键词：自然语言处理；TFIDF；词加权算法；标签提取；监督学习
中图分类号：TP391 **文献标识码**：A

Label Extraction Algorithm Based on Enhanced TFIDF

WANG Jie[1],LI Xujian[2]
(1.*Shandong University of Science and Technology*,*Qingdao* 266590,*China*;
2.*The Key Laboratory of Digital Mine in Shandong*,*Qingdao* 266590,*China*)

Abstract: As a word weighting algorithm,TFIDF plays an important role in natural language processing such as information retrieval and data mining.TFIDF has relatively simple computational model,suitable for large data parallel computation,applied widely in many fields,and with good explanatory characteristics.Based on the above-mentioned characteristics,this paper proposes to amend the weighted results of TFIDF by means of supervised learning based on TFIDF algorithm as well as by introducing weighting factors and word contribution.This algorithm can effectively extract feature labels in natural language processing,and improve the degree of accuracy in this segmentation field.

Keywords:natural language processing;TFIDF;word weighting algorithm;label extraction;supervised learning

1 引言(Introduction)

互联网每分钟都会产生PB级别的信息。如何从这些信息大数据中提取到有用的信息，并结合快速发展并日益成熟的人工智能技术来改善产品是一个迫切需要解决的问题。移动互联网时代，信息所呈现的特征更加个性化、主体化、终端化。数据中存在无限的价值，谁能从海量的信息数据中撅取价值，谁就可以立足于这个数据时代。

20世纪90年代兴起的人工智能科学，成为信息处理相关从业者手中的一把利器。在人工智能技术中，特征提取一直是一个难点，也是一个痛点。有这么一句话在业界广泛流传：数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。那么特征工程到底是什么呢？顾名思义，其本质是一项工程活动，目的是最大限度地从原始数据中提取特征，以供算法和模型使用。这足以说明在人工智能尤其是机器学习中，特征提取是多么重要。

为了解决特征标签提取的问题，本文将介绍在自然语言处理这个具体应用领域中是如何进行特征工程的。为了达到目的，第一步要对语句进行分词[1]。第二步要对完成分词的文章中的每个词进行加权，通过权值的大小来表示词的重要性[2]。在自然语言处理方向中，最著名的词加权技术就是TFIDF。TFIDF(词频逆文本频率)是一种对基于统计的加权方法，用以评估一个字词对于一个文本或者一个语料库的重要程度。TFIDF已经作为一个成熟的算法广泛应用于自然语言处理的各个领域，其中最典型的就是搜索引擎。TFIDF虽然得到了广泛的应用，但是存在一定的不足，尤其是在细分领域，比如关键词提取[3]。

本文提出了一种基于TFIDF的改进加权技术，使TFIDF在自然语言处理的细分领域中的关键词提取应用上达到更好

基金项目：国家重点研发计划课题(课题编号:2017YFC080446).

的效果，通过基础语料库使计算出的权值结果更能表达词对文章的代表程度。

2 TFIDF算法与不足(TFIDF algorithm and its defects)

Salton在1973年提出了TFIDF(Term Frequency&Inverse Documentation Frequency)算法。算法提出后，Salton及其他学者论证了该算法在信息学中的有效性。TFIDF算法主要分为两个部分，分别是词频(TF)和逆文本频率(IDF)[4]。TF是指文档中某个词出现在文章中的频率值越大，则表明该词的重要性越大。逆文本频率(IDF)是指词出现的篇幅越多，其重要性就越低。逆文本频率有效地避免了词的长尾效应[5]，使权值更能准确地表达词的重要程度。TFIDF算法描述为

$$TFIDF=\frac{W_{ij}}{\sum_k *W_{kj}}*\log\frac{|D|}{|\{j:t_i\in d_j\}|}$$

因为TFDIF算法容易理解并且算法复杂度低，可以使用绝大多数的编程语言计算出准确的TFIDF模型。同时TFIDF具有较好的解释性和准确性，这些特性使得TFIDF被广泛地应用，并被应用到自然语言处理和推荐系统领域。但在实践中人们发现TFIDF存在很多的问题，并不能很好地处理所有的应用领域。尤其是在特殊的细分领域中，TFIDF通常表现得差强人意。本文在自然语言处理领域中的标签提取应用中使用改进的TFIDF算法该方法有效地提高了文章标签提取的准确度。

3 文本预处理(Text pre-processing)

对文本进行标签提取，首先要对文本进行预处理。本文所介绍的文本标签提取技术需要进行四个阶段的预处理。通过对文本进行预处理，可以有效地减少算法的运算量，提高结果的精确度。文本预处理的四个步骤分别为：第一步，准备训练集；第二步，对文本进行分词；第三步，将文本使用向量模型表示[6]；第四步，对向量模型进行降维[7]。

本文所介绍的算法是给予监督学习的算法，所以需要准备一个足够丰富的训练集，并且这个训练集需要人为地进行标注主题。在自然语言处理中，语料库是进行监督学习算法的基础，就像人类学习写文章一样，语文老师就像一个庞大且完善的语料库，这个语料库会告诉你每篇文章的类型和中心思想，并监督你学习[8]。本文在进行权重计算时假设已经有一个完善的语料库，有很多不同的主题分类[9]，并且涵盖了所有的分类。每个分类的文章尽可能多地收集到不同风格和不同作者的文章[9]。

在准备好训练集后，需要对每篇文章进行分词[2]。汉语是一种非形态语言，缺乏形态标记，语序和虚词是重要的语法手段。英语语法手段是显性的，并且英语单词之间用空格分割，而中文与英文不同，这给中文分词带来了巨大的困难。目前中科院和Jieba开源项目提供了针对于中文的分词算法，即便如此，对于某些句子的分词还是会扭曲原句的意思[10]，使关键词被拆分成单个汉字。这就需要人为地对特殊句子进行人为的分词。

分词后，所有的文档会形成一个字典。这个字典包括了训练集所有的词汇，词汇被标示成$[(n1,w1)(n2,w2).....(n,w)]$，其中$n_i$表示词的位置，$w_i$表示特定的词语[11]。值得注意的是，词典几乎囊括了所有的汉语词汇和词组，这无疑加大了特征的纬度，所以在预处理的步骤中需要去掉停用词。停用词是指那些出现频率高但是表示意义小的词[12]，比如文本中的数字和助动词“的”，它们大量地出现在文本中，但是却对文章的主题没有任何影响。除了通过专家进行停用词的挑选，在这里同样可以借助于IDF逆文本频率$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|}$进行停用词的判断。通过定义一个阈值，只要$idf_i$超过了阈值，那么这个词就可以看作是是一个停用词，在文本预处理过程中就需要将这些词从词组中剔除。

4 词贡献度(Word contribution)

每篇文章都有自己的主题和中心思想，主题和中心思想可以近似地代表整篇文章。主题和中心思想同时又可以由体现文章主旨的词汇表示，可以由公式$W\approx topic\approx[tag_1.....tag_n]$表示由文章推出标签特征的过程[11]。

在以上前提下可以提出一个叫主题贡献度的概念[13]。所谓的词汇贡献度就是指根据潜在语义分析的概念，将词语放入在不同的主题下的贡献度记做$T_{i\,k}$，那么将一篇文档词袋中的词对文章的贡献度记做

$$T_{1k}*C_{1k}+T_{2k}*C_{2k}+T_{3k}*C_{3k}+.....+T_{nk}*C_{nk}\,,$$

T表示一个词对文档的贡献度，C表示一个词出现在文中的次数。

5 计算词权重(Word weighting calculation)

第二节讲述了如何分词并进行数学表达，第三节讲述了如何进行语料库的设计。本文所介绍的加权算法就是基于以上两节内容的基础。词袋模型只是将分词后的数组按照顺序排列，加权完的词袋模型具有了新的表达形式[(*,*)……(*，*)]，元祖的key代表字典索引值，元组的value代表字典的权值。

$$\frac{W_{i\,j}}{\sum_k W_{k\,j}}*\left(\frac{1+T_i}{1+e^{p_i}}\right)*idf_i\qquad P_i=-\frac{\sum_d P_{d\,i}}{\sum_d P_{d'\,i}}$$

TFIDF作为一个成熟的算法，有着成熟的应用。本文提出的算法在TFIDF的基础之上，目标是更加精确地对词进行加权，$tf_i=\frac{W_{i\,j}}{\sum_k W_{k\,j}}$表示一个词在文本中出现的频率，$idf_i$表示一个逆文档频率，在第二节中的停用词提取就是用的IDF。

$\sum_d D_{i\,d}$表示词i出现在整个语料库中的篇数。

使用$T_{1k}*C_{1k}+T_{2k}*C_{2k}+T_{3k}*C_{3k}+.....+T_{nk}*C_{nk}$求得文档的

总贡献度，在语料库中取出贡献度最高主题T,并求出该主题下词i出现的篇数。$P(i)$表示的是一个词所代表主题的频度，所以$P(i)$是词i在整篇文章出现的次数和词在最高贡献度主题下的出现次数的比值并求负数。例如在一个语料库中秦始皇这个词在历史中出现了100次，在影视中出现了50次，在其余类中总共出现了50词，那么秦始皇这个词$P(i)$分别为−0.5、−0.25、>−0.25、……、$P(i)$虽然能够很好地表示词的主题相关性，但是数值存在差别太大的可能性，因为如果在总数很大的情况下，那么很可能出现$P(i)$的值也过大，计算后的误差也会变得特别大。

所以要对$P(i)$进行归一化。利用逻辑回归函数$\theta(s)=\frac{e^s}{1+e^s}=\frac{1}{1+e^{-s}}$进行归一化，其几何表示如图1所示。

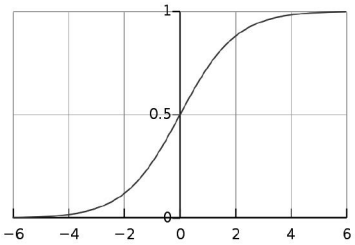

图1 logistic函数几何意义图

Fig.1 The geometric meaning of the logistic

利用逻辑函数的特性，在区间内的值区间为(0,1)，也就是说无论这个数值多么大，它的值区间都很友好。再对其变形并带入$P(i)$和词i的贡献度T_i，经过变换得到$\lambda=\left(\frac{1+T_i}{1+e^{pi}}\right)$。

$\lambda=\left(\frac{1+T_i}{1+e^{pi}}\right)$可以看做是一个影响因子，可以对$TD-IDF$进行修正，这可以叫做词$i$权修正因子。

至此，我们得到了如何加权的算法。根据这个算法我们将算出每个词的权值，并带入元组列表中。

6 结论(Conclusion)

首先通过介绍TFIDF的算法原理以及对TFIDF算法的加权结果的解释可知，这是一个伟大的算法，但随着人工智能和大数据的到来，特征提取变得越发的重要，TFIDF这个在自然语言处理中近乎万金油的算法模型已经不能很好地满足需要，所以在TFIDF算法的基础上进行改进。特征提取是一个复杂的过程，包括多个步骤，每一步都会对结果进行影响，比如分词。好的分词方法可以在分词后不改变原意，让后面的算法可以有效地提取出文本的特征标签。词典和词向量和停用词可以减少模型的时间复杂度和空间复杂度，在监督学习算法的前提下，模型需要大量的数据来学习，面对这些海量的数据，如果前面几步处理的不恰当，很可能导致整个模型的可用性变得很差。

最后针对TFIDF在自然语言处理特征标签提取应用中的不足，对算法进行改正。首先TFIDF体现出自然语言的语义。语义可以说是文本最重要的体现形式。根据TFIDF算法很可能获取的权值较高的特征标签中包括多组反义词，从而导致无效的结果。因为在论证某一问题时不可能避免地会使用它的对立面语义而TFIDF又是忽略语义的，所以引入了词贡献度这个概念可以很好地弥补TFIDF的语义处理上的缺失。最后为了使结果更加平滑，使用逻辑回归函数作为归一化函数。

本文对TFIDF的改进主要在两个方面。一是利用了词贡献度，二是根据词贡献度来得出修正因子，使结果更加准确。词贡献度可以合理针对于主题方面对TFIDF进行了改进，为TFIDF增加影响因子，力图使所得到的权值更加地准确。

参考文献(References)

[1] 韩冬煦,常宝宝.中文分词模型的领域适应性方法[J].计算机学报,2015,38(02):272−281.

[2] 初建崇,刘培玉,王卫玲.Web文档中词语权重计算方法的改进[J].计算机工程与应用,2007,17(19):192−194;198.

[3] 刘勘,周丽红,陈譞.基于关键词的科技文献聚类研究[J].图书情报工作,2012,56(04):6−11.

[4] 施聪莺,徐朝军,杨晓江.TFIDF算法研究综述[J].计算机应用,2009,29(S1):167−170;180.

[5] 陈力丹,霍仟.互联网传播中的长尾理论与小众传播[J].西南民族大学学报(人文社会科学版),2013,34(04):148−152;246.

[6] 江大鹏.基于词向量的短文本分类方法研究[D].浙江大学,2015.

[7] 刘欣,余贤栋,唐永旺,等.基于特征词向量的短文本聚类算法[J].数据采集与处理,2017,32(05):1052−1060.

[8] 刘建伟,刘媛,罗雄麟.半监督学习方法[J].计算机学报,2015,38(08):1592−1617.

[9] 谭金波,李艺,杨晓江.文本自动分类的测评研究进展[J].现代图书情报技术,2005(05):46−49;14.

[10] 莫建文,郑阳,首照宇,等.改进的基于词典的中文分词方法[J].计算机工程与设计,2013,34(05):1802−1807.

[11] 黄栋,徐博,许侃,等.基于词向量和EMD距离的短文本聚类[J/OL].山东大学学报(理学版),2017(07):1−6.

[12] 崔彩霞.停用词的选取对文本分类效果的影响研究[J].太原师范学院学报(自然科学版),2008,7(04):91−93.

[13] 周水庚,关佶红,胡运发.隐含语义索引及其在中文文本处理中的应用研究[J].小型微型计算机系统,2001(02):239−243.

作者简介：

王 杰(1993−)，男，硕士生.研究领域：大数据分析，人工智能，领域驱动设计.

李旭健(1971−)，男，博士，副教授.研究领域：计算机视觉，VR&AR，大数据技术.