关键词自动提取方法的研究与改进

黄 磊[1,2] 伍雁鹏[2] 朱群峰[2]
(湖南大学信息科学与工程学院 长沙 410082)[1] (邵阳学院信息工程系 邵阳 422000)[2]

摘 要 关键词提取技术是信息检索和文本分类领域的基础与关键技术之一。首先分析了TFIDF算法中存在的不足,即IDF(Inverse Document Frequency)权值中没有考虑特征词在类内以及类别间的分布情况。因此,原有的TFIDF方法会出现有些不能代表文档主题的低频词的IDF值很高,而有些能够代表文档主题的高频词的IDF值却很低的情况,这会导致关键词提取不准确。通过增加一个新的权值,即类内离散度DI(Distribution Information)来增加关键的特征词条的权重,提出了一种新的算法DI-TFIDF。实验中使用的是搜狗语料库,选择其中的体育、教育和军事3类文档各1000篇作为实验的语料库,分别用基于传统TFIDF方法和基于DI-TFIDF方法提取关键词。实验结果表明,所提出的DI-TFIDF方法提取关键词的准确度要高于传统的TFIDF算法。

关键词 关键词提取,特征权重,TFIDF,DI-TFIDF
中图法分类号 TP391.1 **文献标识码** A

Research and Improvement of TFIDF Text Feature Weighting Method

HUANG Lei[1,2] WU Yan-peng[2] ZHU Qun-feng[2]
(School of Information Science and Engineering, Hunan University, Changsha 410082, China)[1]
(Department of Electric Engineering, Shaoyang University, Shaoyang 422000, China)[2]

Abstract Keywords extraction method plays a very important role in the areas of text classification and information retrieval. This paper firstly analysed the shortage of the original TFIDF algorithm, that is the IDF (Inverse Document Frequency) algorithm does not consider the distribution of feature term between categories. So some problems will appear, such as the terms with low frequency and the high IDF weights, and some words with high frequency and low IDF weights, which can cause that the precision of keywords extraction is not accurate. After analysis of these problems, by increasing a new weight DI (Distribution Information), we got a new DI-TFIDF algorithm. A corpus used in the experiment was downloaded from the Sogou corpus and we selected the 1000 article of sports, education and military documents as an experiment based on the traditional TFIDF method and the DI-TFIDF method. Experimental results show that our proposed DI-TFIDF method can extract the keywords in a higher accuracy than traditional TFIDF algorithm.

Keywords Keywords extraction, Term-weighting, TFIDF, DI-TFIDF

1 引言

随着Internet的广泛应用,海量的信息资源以文本形式存在。信息世界的不断发展,极大地丰富了人类的生活,但也带来了棘手的问题:如何在庞大的信息世界中迅速找到所需的信息。这一问题成为了一项具有重大研究意义的课题。

在文档信息中,关键词起到了关键作用,它是能够反映一篇文档主题内容的词语或与文档所在领域高度相关的词语,帮助人们在搜寻所需的信息时能够迅速地定位到相应的文档。然而,大量的文档中并没有标注出关键词。人工标注出这些文档的关键词又是非常耗时和困难的,所以迫切需要对关键词进行自动提取。关键词提取技术应运而生,并不断发展,成为了网页浏览、文本分类和信息检索等领域的技术基础,帮助人们迅速找到相应的文本信息,满足了人们对信息需求的渴望。

综上所述,关键词提取技术是文字信息处理中重要的基础工作。本课题研究的目的是基于改进的TFIDF算法提取出关键词,由于文本特征权重算法对关键词提取的准确率有着重要的影响,因此对传统的TFIDF的改进就非常有必要。最终研究成果是设计出关键词提取系统,该系统可以应用到网页关键词提取、文本分类和信息检索系统上。研究这样的关键词提取系统,可以在一定程度上帮助用户更为准确和快速地搜寻到相应的信息,有利于信息的传播和知识的推广,并减轻人工标注关键词的负担,具有深刻的意义。

2 国内外研究现状和成果

国外对关键词提取研究较早,20世纪50年代,美国的Luhn[1]进行了开创性的研究,提出了基于词频统计的抽词标

到稿日期:2013-11-20 返修日期:2014-03-18 本文受湖南省教育厅一般项目(09C887):基于语义网的网络教学资源检索系统研究资助。
黄 磊(1976—),女,硕士,主要研究方向为信息处理、语义Web技术;伍雁鹏(1975—),男,博士生,副教授,主要研究方向为计算机应用;朱群峰(1974—),女,硕士,主要研究方向为计算机控制。

引法。经历了50多年的发展，对自动标引的研究逐渐消失，其原因是传统的自动标引方法的效率达到了极限，人们广泛地使用全文索引，且索引功能能满足用户的需求。在1963年，美国的Chemical Abstracts[2]开始用电子计算机编制关键词索引，提供了快速检索文献资料主题的途径。这也是统计分析方法最早被应用到关键词自动提取。20世纪70年代初，Lois[3]开始采用句法分析等语言学习方法与词频统计方法相结合的方法来提取关键词。20世纪90年代末到现在，关键词提取的研究也逐渐深入，许多学者提出了不同的方法，取得了令人瞩目的成绩，如90年代末Turney[4]提出了基于遗传算法GenEx的关键词提取方法，Witten[5]提出了基于朴素贝叶斯的关键词提取方法。近年来，关键词提取的研究趋于活跃，2003年Tomokiyo与Hust[6]提出了基于语音模型的关键词提取方法，Hulth[7]利用Bagging算法进行了基于集成学习的关键词提取。2006年Samhaa[8]提出以标点符号和停用词为词语间隔，先提取出一个词语序列，再以此序列和序列的N-gram为候选对象，计算候选关键词的特征项的TFIDF、位置、短语长度等特征值，进而从候选关键词中提取特征值大的作为关键词。2007年，Ercan和Cicekli[9]提出基于TFIDF改进的，通过词汇链来增强词语之间语义联系的方法。2008年Niraj[10]通过LZ78压缩算法获取N-gram，然后简单地过滤出不合适的词语，计算这些词语的权重，最后提取权重大的作为关键词。

国外对TFIDF的研究较早，提出的改进方法主要有：1999年Roberto Basils[11]提出了TF * IWF * IWF，该改进方法有效提高了特征词在语料库的权重，但没有充分考虑到词的重要性，因为特征权重并不仅仅是由词语在语料库中出现的频率决定，而是由词语在文档和语料库中出现的频率共同决定，这使得该算法还存在着不足。2004年Bong Chih How和Narayanan[12]根据不同类别的文档数可能存在数量级的差距，提出利用CTD（Category Term Descriptor）来改进TFIDF，以改善类别数据集偏斜所引起的误差。

国内也有很多学者对TFIDF算法进行了研究，并且取得了显著的成果。2006年张玉芳等人[13]通过修改IDF的计算方法，增加那些在一个类中频繁出现的特征项权重，考虑了IDF在类别中的分布情况，有效提高了准确率。2007年张玉芳、陈小莉等人[14]把信息论中的信息增益应用到文本集合的类别层次上，提出了一种改进的TF * IDF * IG算法。2008年沈志斌和白清源[15]提出了BOR-TFIDF（Based On Ratio-TFIDF）算法，该算法重新针对特征词对类别的区分度进行了调整。文献[14,15]均考虑不同类别之间的分布状况，但未能深入到类别内部，没有考察类内的分布情况。文献[16]阐述了TFIDF算法在不同领域的改进，学者们在组合型歧义切分字段、聊天文本权重计算、网页权重等方面提出了不同的改进方法，提高了TFIDF算法对不同领域文本的处理能力。2011年张保富等人[17]提出了一种结合特征项的类间和类内信息分布熵的TFIDF特征加权方法。该方法对类间分布熵和类内分布熵进行了分析，综合考虑了特征项在类间和类内的分布情况对其在文本中的贡献的影响，弥补了传统TFIDF算法的不足。2012年李学明等人[18-21]提出了基于信息增益与信息熵的TFIDF算法，该方法针对信息增益只考虑特征词在类间的分布情况，而没有考虑特征词在类内的分布情况的问题，将能表示特征项在类内分布程度的信息与信息增益综合起来考虑，利用信息熵对特征词权重进行调整，从而提高了特征词权重的计算精度，提高了关键词提取的准确度。

根据以上对国内外的研究成果的分析，可得出语言分析对词典的依赖较大，提取效果有赖于词典的完整性。而人工智能也同样需要训练库和知识库，对它们的依赖较大，机器学习的能力影响了提取的效果。因此本文重点研究TFIDF算法，通过介绍和研究传统的TFIDF算法，发现其存在的不足并有针对性地加以改进，提出了新的DI-TFIDF算法。

3 特征权重算法TFIDF的改进

文本是由词语构成的，要在文本中提取出关键词，就必须赋予特征项相应的权重，权重越大的特征项越能代表文本的主题。特征权重算法TFIDF中TF可以反映特征项在文本中出现的频率，IDF可以反映特征项反比于文档集中出现特征项的频率，很好地结合了TF和IDF。

3.1 TFIDF

TFIDF实际上是TF * IDF，TF是指特征项在文档中出现的次数，IDF是指反文档频率。其计算公式是：

$$w_{ij}=tf_{ij}\times idf_i=tf_{ij}\times\log(\frac{N}{n_i}+0.01)$$

由于考虑到文档的内容长度会影响到权值计算，对上式进行归一化处理，得到如下公式：

$$w_{ij}=\frac{tf_{ij}\times\log(\frac{N}{n_i}+0.01)}{\sqrt{\sum_{i=1}^{N}tf_{ij}^2\times\log^2(\frac{N}{n_i}+0.01)}}$$

其中，tf_{ij}是特征项t_i在文档d_j中出现的次数，idf_i是指出现特征项t_i的倒数，N为文档集的文档总数，n_i是出现特征项t_i的文档数。

3.2 TFIDF的不足

用传统的TFIDF公式来提取关键词，一般存在两个问题：1）有些不能代表文本的低频词，IDF的值反而很高；2）有些能够很好代表文本的高频词，IDF值却很低。主要的原因是TFIDF中IDF的算法没有考虑到特征项在类间和类内的分布情况。具体分析如下：

（1）IDF没有考虑到特征项在类间的分布信息

假设某一类的C_i包含词条t_i的文档数为n，而其他类包含的t_i的文档数为m，包含词条t_i的文档数为w，$w=n+m$，w随着n变化，n增大时w也变大。可是按照IDF的公式，得到的IDF值却很小。但是按实际分析可知，n值大就表示词条t_i在C_i类中频繁出现，可以作为特征词代表这一类，应该赋予较高的权重。另一方面，如果包含词条t_i的文档总数w小，而词条t_i均匀地分布在各个类间，则该词条t_i不具有区分能力，不应该作为特征词，应该赋予较低的权重。但是按照传统的TFIDF算法，得出的IDF值却很大，与分析的结果相反。

（2）IDF没有考虑到特征项在类内的分布信息

在同一类中不同的特征项，类内分布均匀的应该比分布不均匀的特征权重要高，但是按照传统的TFIDF算法，IDF值却很大。假设有3个类别C_1，C_2，C_3，每个类别各3篇文档。分析t_1，t_2，t_3，t_4 4个特征项在各个类别和文档中的分布

情况，如表 1 所列。

表 1　各个特征项在文档中出现的频率

文档/特征项	C_1			C_2			C_3		
	1	2	3	1	2	3	1	2	3
t_1	5	4	6	0	0	0	0	0	0
t_2	4	0	0	0	6	0	0	0	10
t_3	3	0	4	0	0	2	0	0	0
t_4	1	1	1	1	1	1	1	1	1

用传统 TFIDF 算法计算各个特征项的权值(没有进行归一化)，如表 2 所列。

表 2　各个特征项的 TFIDF 值

特征项	C_1			C_2			C_3		
	1	2	3	1	2	3	1	2	3
t_1	2.39	1.91	2.87	0	0	0	0	0	0
t_2	1.91	0	0	0	2.87	0	0	0	4.79
t_3	1.43	0	1.91	0	0	0.96	0	0	0
t_4	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004

从表 2 中可以看出：特征项 t_4 在 3 个类中均匀分布，表明其没有关键信息，所以权值很低，这是 TFIDF 算法的优点，即能够过滤掉均匀分布的特征项。t_1 只在 C_1 中出现，区别能力最强。而 t_2 在 3 个类中均出现，区别的能力也最弱。但从表 2 中可以发现 t_2 权值却很高，这是因为根据 TFIDF 的算法，如果 IDF 相同，TF 就决定了特征项的权重的大小，而文档集中含有特征项 t_1、t_2 的文档数相同，含有 t_2 的文档中 TF 较高，因而产生了不合理的结果。这些可能由于偶然因素导致的特殊词条，其权重应该较小，但是如果按照传统的 TFIDF 算法计算，却会得到较高的权重。这就是传统 TFIDF 算法中 IDF 没有考虑特征项在类间、类内的分布情况而产生的误差。

3.3　改进的 TFIDF

针对 TFIDF 的不足，提出改进的意见。根据第一个问题，过滤掉低频词和只出现在一个句子中的词语和单字，来改善效果。根据第二个问题，本文从类间、类内离散度出发，提出基于特征项分布差异的 DI-TFIDF 特征权重改进算法。

(1)由于 IDF 没有考虑到特征项在类间的分布信息，我们考虑对 IDF 加以改进，增加那些在一个类中频繁出现的特征项的权重。改进的 IDF 算法为：

$$IDF=\log(\frac{n_i}{n_i+m}N+0.01)$$

其中，总文档文本数为 N，包含特征词条 t_i 的文档数为 w，$w=n_i+m$。n_i 是其中某一类 C_i 中含有特征词条 t_i 的文档数。m 表示文档集中其他类含有特征词条 t_i 的文档数。设 $f=\frac{n_i}{n_i+m}=\frac{1}{1+\frac{m}{n_i}}$。

当含特征词条 t_i 的文档数 w 一定时，n_i 越大，则 m 越小、f 越大，因而 IDF 越大、TFIDF 也越大。也即如果某一个类 C_i 中包含特征词条 t_i 的文档数多，而其他类中包含特征词条 t_i 的文档数少，则 t_i 能够代表这个类 C_i 的特征，应赋予较大的权重。故改进的算法能够有效弥补传统 TFIDF 算法没有考虑类间分布信息的不足。

(2)由于 IDF 没有考虑到特征项在类内的分布信息，我们考虑增加类内离散度 DI 来观察特征项在类内的分布情况。离散度是各个文档中特征词的差异程度，可以很好地反映在同一个类中不同文档特征词频率的不同。类内的离散度计算公式如下：

$$DI=\frac{\sqrt{\frac{[\sum_{j=1}^{n}(tf_{ij}-\overline{tf_{ij}})^2]}{(k-1)}}}{tf_{ij}}$$

其中，$\overline{tf_{ij}}=\frac{1}{n}\sum_{j=1}^{n}tf_{ij}$；$k$ 为类内总的文档数；tf_{ij} 表示特征词 t_i 在第 j 篇中出现的次数；$\overline{tf_{ij}}$ 是特征词 t_i 在类内各个文档中出现次数的平均值。如果特征词只在某一篇文档中出现，说明其分类能力差，类内的离散度 DI 可以取得最大值 1；如果特征词在类内文档中每篇文档的 TF 值都相同，则说明其分类能力好，类内离散度 DI 可以取到最小值 0。

在传统 TFIDF 算法的基础上，我们对 IDF 进行改进并增加类内离散度，最终得到 DI-TFIDF 算法，算法的公式如下：

$$w_{ij}=tf_{ij}\times\log(\frac{n_i}{n_i+m}N+0.01)\times(1-DI)$$

考虑到类内离散度与特征词的分类能力成反比，在构造特征权重计算公式时用$(1-DI)$表示。对上式进行归一化处理得到公式：

$$w_{ij}=\frac{tf_{ij}\times\log(\frac{n_i}{n_i+m}N+0.01)\times(1-DI_{(t_i,c)})}{\sqrt{\sum_{i=1}^{N}tf_{ij}^2\times\log^2(\frac{n_i}{n_i+m}N+0.01)\times(1-DI_{(t_i,c)})^2}}$$

其中，w_{ij} 表示特征项 t_i 在某类别 C 中 j 篇文档的权重，tf_{ij} 表示特征项 t_i 在文档中出现的次数，$IDF=\log(\frac{n_i}{n_i+m}N+0.01)$表示特征项 t_i 的反文档频率，$DI_{(t_i,c)}$ 表示特征项在类别 C 中 j 篇的 DI 值。我们对表 1 中各个特征项在类别中的 DI-TFIDF 权重进行计算(没有进行归一化)，计算结果如表 3 所列。

表 3　各个特征项的 DI-TFIDF 值

特征项	C_1			C_2			C_3		
	1	2	3	1	2	3	1	2	3
t_1	4.77	3.14	5.05	0	0	0	0	0	0
t_2	1.01	0	0	0	1.52	0	0	0	2.53
t_3	2.06	0	2.84	0	0	0.82	0	0	0
t_4	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004

通过计算结果可以看出，t_2 在各个类别的特征项权重比 t_1 在各个类别的特征项权重小，说明 DI-TFIDF 算法有效解决了文档在类间、类内的分布情况，使得到的权重更为准确。

4　实验及结果分析

从互联网搜狗语料库中选取体育、教育、军事 3 类各 1000 篇文档作为实验所需语料库。其中训练样本和测试样本分布都有 3 类，训练样本共有 3000 篇文档，测试样本有 150 篇文档，在各个类中，训练文档和测试文档的比例是 20∶1。为了验证改进算法的有效性，本文进行了两组实验，分别为基于传统 TFIDF 的关键词提取和基于 DI-TFIDF 的关键词提取，采用查全率、查准率对提取的结果进行评价。基于传统 TFIDF 算法和 DI-TFIDF 算法的关键词提取效果如表 4—表 6 所列。

表4 基于TFIDF算法(a)和DI-TFIDF算法(b)的实验结果(体育类)

(a)

特征维数	查全率	查准率
500	66.3	56.3
1000	68.5	57.5
2000	70.8	58.8
4000	71.0	59.6
6000	72.7	61.6
8000	74.3	62.3
平均值	70.6	59.4

(b)

特征维数	查全率	查准率
500	72.1	61.9
1000	73	63.5
2000	74.5	64.5
4000	75.6	66.7
6000	76.4	67.6
8000	77.5	68.1
平均值	74.9	65.3

表5 基于TFIDF算法(a)和DI-TFIDF算法(b)的实验结果(军事类)

(a)

特征维数	查全率	查准率
500	54.1	46.6
1000	55.6	47.5
2000	56.6	48.9
4000	57.8	49.6
6000	58.4	51.1
8000	59.4	52.1
平均值	57	49.3

(b)

特征维数	查全率	查准率
500	63.2	51.9
1000	65.5	52.6
2000	66.7	53.8
4000	67.3	55.3
6000	68.4	56.7
8000	69.5	58.2
平均值	66.8	54.8

表6 基于TFIDF算法(a)和DI-TFIDF算法(b)的实验结果(教育类)

(a)

特征维数	查全率	查准率
500	58.3	51.3
1000	59.9	52.6
2000	60.6	53.5
4000	61.2	55.6
6000	62.3	56.7
8000	63.5	57.2
平均值	61	54.5

(b)

特征维数	查全率	查准率
500	63.8	54.3
1000	64	55.7
2000	65.1	57.9
4000	67.2	58.1
6000	68.1	59.4
8000	69.9	60.6
平均值	66.4	57.7

从以上数据可以看出，改进的DI-TFIDF方法提取关键词的效果比传统的TFIDF方法提取关键词的效果要好，在查全率、查准率上都有了一定的提高。因为DI-TFIDF算法考虑了特征项在类间和类内的分布情况，对于那些在某个类别中频繁出现的特征项赋予了较高的权重，降低了在类内文档中偶然出现的特征项的权重。因此DI-TFIDF算法对正确提取关键词起到了一定的积极作用。

结束语 本文以关键词提取为研究对象，对现有的关键词提取进行了总结，介绍了国内外对关键词提取的研究成果，并对关键词提取中具有重要影响的特征权重TFIDF算法的发展及国内外TFIDF改进成果进行了介绍。对特征权重算法做了详细的研究，提出了改进的方法：DI-TFIDF算法，并分别对基于传统的TFIDF算法的关键词提取和基于DI-TFIDF算法的关键词提取进行了实验，结果表明基于DI-TFIDF算法的关键词提取效果要比传统算法好，证实了改进的有效性。

但是由于时间和能力的限制，还有很多的工作需要改进和深入。有几个方面需要在以后的研究中进行改进：

(1)对中文分词的研究较少，而是引用了现有的分词工具，根据实验分词效果还不够理想，下一步需要研究采用分词效果更好的分词工具。

(2)对TFIDF算法的改进，仅考虑了特征词在类间、类内的分布情况，并未考虑特征词的词性、特征词长度和特征词在文档中出现的位置，导致了特征权重计算不够准确，需要在以后工作中不断研究和测试数据，并根据现有的改进方法提出更有效的改进方法。

参 考 文 献

[1] Luhn H P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J]. IBM Journal of Research and Development, 1957, 1(4): 309-317

[2] Edmundson H P, Oswald V A. Automatic Indexing and Abstracting of the Contents of Documents[R]. Planing Reserarch Corp, Document PRC R-126, ASTLA AD No. 231606. Los Angeles, 1959: 1-142

[3] Lois L E. Experiments in Automatic Indexing and Extracting [J]. Information Storage and Retrieval, 1970, 6: 313-334

[4] Turney P D. Learning to Extract Keyphrases from Tex [R]. NRC Technical Report ERB-1057. National Research Council, Canada, 1999: 1-43

[5] Witten I H, Paynter G W, Frank E, et al. Practical Automatic Keyphrase Extraction[C]// California: Proceedings of The 4th ACM Conference on Digital Libraries. 1999: 254-256

[6] Tomokiyo T, Hurst M. A language Model Approach to Keyphrase Extraction[C]// Proceedings of the ACL Workshop on Multiword Expressions: Ananlysis, Acquisition & Treatment. Sapporo, Japan, 2003: 33-40

[7] Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge[C]// Proceeding of the 2003 Conference on Emprical Methods in Natural Language Processing. Sapporo, Japan, 2003: 216-223

[8] Samhaa R. A Simple System for Effective Keyphrase Extraction [C]// Proceeding of 3th IEEE International Conference on Innovations in Information Technology. 2006: 1-5

[9] Ercan G, Cicekli I. Using Lexical Chains for Keyword Extraction [J]. Information Processing & Management, 2007, 43(6): 1705-1714

[10] Niraj K, Kannan S. Automatic Keyphrase Extraction from Scientific Documents Using N-Gram Filtration Technique[C]// Proceeding of DocEng'08 Conference. 2008: 199-208

[11] Basils R, Moschitti A, Pazienza M. A text classifier based on linguistic processing[C]// Proceedings of UCAI, Machine Learning for Information Filtering. 1999: 36-40

[12] How B C, Narayanan K. An empirical study of feature selection for text categorization based on term weightage[C]// Proceeding of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. Washington DC: IEEE Computer Society, 2004: 599-602

[13] 张玉芳，彭时名，吕佳. 基于文本分类TFIDF方法的改进与应用[J]. 计算机工程，2006，32(19)：77-78

[14] 张玉芳，陈小莉，熊忠阳. 基于信息增益的特征词权重调整算法研究[J]. 计算机工程与应用，2007，43(35)：159-160

[15] 沈志斌，白清源. 文本分类中特征权重算法的改进[J]. 南京师范大学学报：工程技术版，2008，8(4)：95-149

[16] 施聪莺，徐朝军，杨晓江. TFIDF算法研究综述[J]. 计算机应用，2009，29(6)：167-170

[17] 张保富，施化吉，马素琴. 基于TFIDF文本特征加权方法的改进研究[J]. 计算机应用与软件，2011，28(2)：17-21

[18] 李学明，李海瑞，薛亮，等. 基于信息增益与信息熵的TFIDF算法[J]. 计算机工程，2012，37(8)：37-40

[19] Wang D X, Gao X, Andreae P. Automatic Keyword Extraction from Single-Sentence Natural Language Queries[C]// PRICAI 2012. Berlin: Springer-Verlag, 2012: 637-648

[20] 张颖颖，谢强，丁秋林. 基于同义词链的中文关键词提取算法[J]. 计算机工程，2010，36(19)：93-95

[21] 刘铭，王晓龙，刘远超. 基于词汇链的关键短语抽取法的研究[J]. 计算机学报，2010，33(7)：1246-1255