科技项目查重中特征词 TF-IDF 值计算方法的改进 *

方延风

(福建省科学技术信息研究所　福州　350003)

摘　要:针对科技项目查重的需要,利用分词的结果,将科技项目文档转换为文本向量空间模型,抽取特征词,并将特征词的位置和词的长度 2 个因素考虑进来,提出一种 TF-IDF 值的改进计算方法,并实例验证该方法有一定的改善效果。

关键词:文本特征　特征词权值　TF-IDF　算法改进　项目查重　科技项目

中图分类号:TP301.6　　文献标识码:A　　doi:10.3969/j.issn.1005-8095.2012.01.001

使用空间向量模型方法对科技项目进行查重,一般要先对文本进行分词,但如果直接用分词的结果作为文本向量中的各个维,则整个文本向量空间的维数将非常高,这会降低整个处理过程的效率,且会损害聚类算法的精确性。因此,为了降低向量空间维数,简化计算,提高文本处理效率,需在不损害文本核心信息的情况下尽量减少要处理的单词数。一般的方法是根据某个特征评估函数来计算各个词的特征得分,又称为权值,然后按分值对这些特征进行排序,选取若干个分值最高的作为特征词。这些特征词即可作为文本的中间表示形式,用来实现文本之间的相似度计算、文本聚类等。特征提取算法的优劣将直接影响到系统的运行效果,而决定文本特征提取效果的主要因素是评估函数的质量。

当前有多种文本特征的评估算法,如 TF-IDF、词频方法、互信息、期望交叉熵、二次信息熵、信息增益方法、x^2 统计方法、遗传算法、主成分分析法、模拟退火算法、N-Gram 算法等。这些评估算法在不同领域中得到使用,也各自存在缺点和不足。其中 TF-IDF 因为计算较为简单并且效果良好,应用十分广泛。

1　影响特征词权值的因素

1.1　词频

词频是特征提取中必须考虑的重要因素,一般认为文本中的中频或低频词汇更能反映文本特征,而高频词则作用不大。极端的例子如"的"这样的词,在几乎所有文本中都频繁出现,但对文本特征贡献为零,所以一些算法中直接用停用词表将此类高频词去除。

张鹏飞等[1]指出:大规模文本的词频统计显示,每个词在常用文本中均有相对固定的一个词频范围和篇章覆盖程度,特定的词在特定类别文本中的词频和篇章覆盖程度也是相对固定的,而且这个特征与该词在其他类别中的特征也是存在差别的。如某个特征词 X 在一般文本中属于中频词,在类别 $T1$ 中变成了高频词,而在 $T2$ 中变成了低频词,那么这个词可以被用来作为区分 $T1$、$T2$ 及其他类别的特征词。

1.2　词性

中文文本中能标识文本特性的一般是实词,如名词、动词、形容词等。而感叹词、介词、连词等虚词对确定文本类别基本没有意义,如果将虚词作为文本特征词,会带来很大噪音,降低文本分类的效率和准确率。因此,提取文本特征时应去除虚词,只提取文本中的名词和动词等实词作为文本的特征词。

1.3　位置

一般文本的标题简练、醒目,与文本的主要内容有着重要的联系,通常是文本中心内容的高度概括。文本中类似"总之"、"综上所述"等概括性语义后的句子,往往也包含了文本的中心内容。因此,这些位置中的词应该具有更加重要的地位。李爱仕[2]提到,美国的 P.E.Baxendale 的调查结果显示:段落的论题是段落首句的概率为 85%。因此,我们可以认为,首段、末段、段首、段尾、标题和副标题、子标题、概括句等处的句子在较大程度上概述了文章的内容,对于这些位置句子中的词,其相应表示文本内容或区别文本类别的能力较高,所以在单词权值计算中应考虑该词的位置信息,对上述位置的词增加权重。

1.4　专业词库

一般通用词库里的词大多不会成为科技项目文本中的特征词,为了提高系统运行效率,应根据科技项目文本的特点建立专业的分词词典,通过人工确定领域内的关键词集,这样在保证特征提取准确性的前提下,可提高系统的运行效率。

1.5　词的长度

一般中文中较长的词往往反映比较具体、下位的概念,而较短的词则表示相对抽象、上位的概念。长词的频率较低,是面向内容的;短词具有较高的频率和更多的含义,是面向功能的。因此,增加长词的权重,有利于对词汇进行分割,从而更准确地反映出

收稿日期:2011-10-13

* 本文系福建省公益类科研院所科研专项"基于相似性计算的科技项目查重技术研究"(项目编号:2010R1009-4)的成果之一。

作者简介:方延风(1975—),男,硕士,高级工程师。

特征词在文章中的重要程度，尤其在科技项目的文档中的关键词多是一些专业学术组合词汇，长度较长，其含义更明确，更能反映文本主题。

1.6　同义归并

向量空间模型最基本的假设是各个分量间正交，但实际上，作为分量的词汇间可能存在很大的相关性，模型的假设是无法满足的。因此，还需要对文本进行部分语义分析，如进行同义词合并等。科技项目文本中，经常可见“计算机”、“电脑”等含义相同或者近似的词，应该将它们合并成同一词。这样得到的词作为文档向量的特征项，可以减少特征之间的相关性，降低文档向量的维数，提高特征提取的精度和效率。但目前这方面的算法比较复杂，本文在这方面不做进一步的讨论。

2　常用的 TF-IDF 值算法

2.1　TF-IDF 值算法

TF-IDF(Term Frequency-Inverse Document Frequency)是一种统计方法，用以评估一个词对于一个文件集的重要程度，其中 TF 称为词频，用于计算该词描述文档内容的能力，IDF 称为反文档频率，用于计算该词区分文档的能力。TF-IDF 基于以下设定：一个词在一个文本中出现很多次，则在另一个同类文本中出现多次的概率也很大，反之亦然。因此在特征空间坐标系中取 TF 词频作为测度可以体现同类文本的特点。考虑到单词区别不同类别的能力，TF-IDF 引入了逆文本频度 IDF 的概念，认为一个单词出现的文本频数越小，它区别不同类别文本的能力就越大。因此，可以用 TF 和 IDF 的乘积作为特征空间坐标系的取值测度，并用它完成对权值 TF 的调整，以突出重要单词，抑制次要单词。

TF-IDF 有多种计算公式，一种较为常用的计算公式如下：

$$TF_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

其中，$n_{i,j}$ 是该词在文件 d_j 中的出现次数，而分母则是在文件 d_j 中所有字词的出现次数之和。

$$IDF_i=\log\frac{|D|}{|\{d:d\ni t_i\}|+1} \tag{2}$$

其中，$|D|$ 为文件总数；$|\{d:d\ni t_i\}|$ 为包含词语 t_i 的文件数目。

$$TFIDF_{i,j}=TF_{i,j}\times IDF_i \tag{3}$$

2.2　据位定权

在使用 TF-IDF 计算特征词的权重时，并没有考虑特征词在文本中的位置、特征词的长度这两个重要信息。有调查显示，段落首句揭示段落主题的概率超过 80%，所以若特征词在首句出现，其重要性应该适度提高。因此，单纯由 TF-IDF 计算出的权重并不能完全准确地反映出特征词在文章中的重要程度，在设计权重评估函数时，要考虑特征词的位置及词的长度等因素。为此，王海涌等[3]提出，设置特征词所处的位置标记 b，特征词的长度 L 及偏移量 a，给出新的评价函数：

$$\Psi=b+[1-\frac{1}{(1+a)\times TF}]\times[1-\frac{1}{L}]\times[1-\frac{DF}{N}] \tag{4}$$

其中，位置标记 b 的取值做如下规定：若该词在标题、副标题中，则 b=2，若为一般特征词，在段首或段尾，则 b=1，其他则 b=0；a 一般取 0.05；DF 为含有 t_i 的文档数目；N 为所有文档的数目。

3　TF-IDF 值算法的改进

综合前文所述的计算方法，本文将特征词的位置(仅考虑出现在首句的情况)和词的长度两个因素考虑进来，对 TF-IDF 值计算方法进行改进。

3.1　候选词的确定

目前大多数中文分词工具的准确率相当高，但对专业术语的识别率还不够高，倾向于将词切分为多个上位词。为了提高关键词抽取的准确率，在科技查重中，需要引入专业科技词库以保证分词的质量。

文本分词的结果中存在大量“的”、“地”、“得”之类的助词，以及“但是”、“因此”等只能反映句子语法结构的词语，它们不能反映文本的主题，而且还会在特征词抽取时产生干扰，应该进行删除。根据科技项目申报的文本材料特点，我们认为词性为虚词的分词对表现主题几乎没有影响，可以删除。

项目名称是对整个项目的高度概括，具有非常重要的作用，本文将项目名称插入成为项目文本的首句，所以文本第一句的分词结果中的实词，必须得到更高的权重。我们设定文本首句(即科技项目的名称)的位置权重为 1.8[4]。其他位置的判断比较复杂，大幅增加了程序的复杂度，而效果却不十分明显，因此本文暂不考虑其他位置权重。

3.2　特征词权重计算

本文在使用 TF-IDF 计算特征词的权重时，考虑了项目名称的重要地位，特征词在项目名称中出现，其重要性显然应该更高。为了减少程序的复杂性，本文将项目名称的字符串直接插到项目文本的第一句，然后加大出现在文本首句的特征词权重。一般而言，较长的词相比较短的词包含了更多信息，更适合表征文本的分类，也需要增大较长词的权重。因此本文提出一种改进后的 TF-IDF 值计算公式：

$$\omega=(1-\frac{1}{l_i+k_1})\times(1+k_2)\times TF_{i,j}\times IDF_i \tag{5}$$

其中，k_1 为常数，本文中取 k_1=2，当特征词在第一句（本文将项目名称直接作为项目文本的首句）时，k_2=0.8，其他时候 k_2=0；l_i 为词特征词的长度。

进行以上计算需要 4 个步骤：

基于灰色 BP 神经网络的 Web 组合服务选择研究 *

廖辰瀚[1]　郑建国[1]　宋　华[2]

(1 东华大学旭日管理学院　上海　201100)(2 安徽财经大学信息工程学院　蚌埠　233041)

摘　要:提出灰色 BP 神经网络 Web 组合服务选择模型,并通过 MATLAB 仿真实验,验证此模型能较好地处理不确定性情况下的 Web 组合服务选择问题,获得全局服务质量(QoS)最优的组合服务方案,明显提高了组合服务 QoS 指标的预测精度。

关键词:Web 组合服务　服务质量　GM(1,1)模型　神经网络　灰关联度

中图分类号:TP389.1　　文献标识码:A　　doi:10.3969/j.issn.1005-8095.2012.01.002

1　引言

随着 Web 服务的发展,单个 Web 服务已经无法满足用户的需求,Web 组合服务应运而生,且相同功能的 Web 组合服务增加迅速。但是这些 Web 组合服务的服务质量(Quality of Service,简称 QoS)相差很大。如何从大量的具有相同功能的 Web 组合服务中,选择出一组服务,使得所组合出的服务具有最好的质量、最高的用户满意度成为 Web 组合服务研究的热点。

在全局优化的 QoS 选择方法的研究中,基于路径拆分的方法[1]将复杂路径的 Web 组合服务拆分为多个相对简单的路径的组合服务,将服务选择问题转换为整数规划问题进行服务的选择工作,但满足一定约束条件下的服务选择问题最终证明是 NP(Non-deterministic Polynomial)难问题[2]。为此,人们提出了许多方案来解决基于 QoS 的 Web 组合服务选择问题,其中基于遗传算法的解决方案[3]是一种新颖的全局优化解决方案,但针对海量 QoS 数据,这些方案算法的运行效率不高,而且结果往往局部最优而不是全局最优。D. Chakraborty 等[4]根据 QoS 指标不确定性特点,建立了基于指标依赖的单目标满足、多目标优化的 QoS 评价模型,此模型以 QoS 指

(1)统计文本集里独立词的个数,并分别计算它们的 IDF;(2)对每个文本统计每一个独立词出现的次数,计算 TF 值;(3)用独立词的 IDF 与每个文本中独立词的 TF 进行计算,得到每个文档中每个独立词的 TF-IDF;(4)根据每个独立词的情况,给出加权的系数,得到最后的 ω 值。

4　TF-IDF 值改进算法的实例分析

某个科技项目的名称为:“三维模型脚型驱动的个性化鞋楦定制 CAD 关键技术研究”,经过公式(3)和公式(5)两种方法计算后得到 TF-IDF 值最大的 10 个特征值以及进行加权后的前 10 个特征词如表 1 所示 (表格中深色底纹的特征词为在项目标题中出现的特征词)。

表 1　某项目特征词权值加权前后比较

序号	特征词	TF-IDF	加权后的特征词	加权后的 TF-IDF
1	楦型	0.142 608	三维模型	0.169 565(加权 1.5)
2	鞋楦	0.117 392	鞋楦	0.158 479(加权 1.35)
3	实体模型	0.114 123	楦型	0.142 608
4	三维模型	0.113 043	实体模型	0.114 123
5	路径	0.104 349	路径	0.104 349
6	轨迹	0.060 870	轨迹	0.060 870
7	参数值	0.052 173	参数值	0.052 173
8	坐标	0.050 349	坐标	0.050 349
9	解调	0.047 827	解调	0.047 827
10	光栅	0.013 043	光栅	0.013 043

从直观的文字理解上看,这个项目标题中的“三维模型”、“鞋楦”这两个词较为明显地体现了项目的内容特征,经过加权计算,这两个词的权重上升到了第一和第二的位置,比原计算得到的 TF-IDF 值更具有合理性。因此,我们认为本文所使用的加权公式(5)是有一定的改善效果。

5　结语

引入 IDF 可以消除一个文档内无意义高频词的影响,但是也带来了一些问题,如一些偶然出现的低频词的权值计算值比较大,可能在事实上放大了这些低频词的重要性。另外,一些热点词汇并非无意义,却因为在大量文档中出现而降低了重要性,比如“网络”这个词,不仅仅出现在信息类、软件类的科技项目文档中,各行各业的项目都可能出现这个词,它的 IDF 值较低,而实际上它在很大程度上还是可以体现一个项目的内容特点的。本文采用的权值计算方法在上述两个方面存在着相应的缺点。

参考文献

[1]　张鹏飞,等. 基于相对词频的文本特征抽取方法[J]. 计算机应用研究,2005(4):23-26

[2]　李爱红. 试论自动摘要技术 [J]. 图书情报工作,2000(4):40-42

[3]　王海涌,等. 基于文本表示的特征项权值确定方法研究[J]. 甘肃科学学报,2005(117):86-89

[4]　刘海峰,等. 文本分类中基于位置和类别信息的一种特征降维方法 [J]. 计算机应用研究,2008,25 (8):2292-2294