数字图书馆文本分类特征提取方法研究与改进

鲍凌云
(菏泽学院图书馆，山东 菏泽 274015)

摘 要：从数字图书馆应用文本分类的重要性入手，介绍了文本分类的含义及基本技术，重点分析了文本分类中常用的两种特征提取方法：互信息算法和 χ^2 统计量算法，指出两种算法存在的不足并提出相应的改进措施。

关键词：文本分类；特征提取；互信息算法；χ^2 统计量算法

中图分类号：G250 **文献标识码**：A **文章编号**：1002-1248（2014）07-0033-03

Research and Improvement of Feature Selection for Page Categorization in Digital Library

BAO Ling-yun
(Library, Heze College, Heze 274015, China)

Abstract: Firstly this paper analyzed the importance of applying page categorization in digital library, then it introduced the connation and basic technologies of page categorization. Also the author studied two main feature selection methods specially: Mutual Information algorithm and Chi-square algorithm. Meanwhile, the paper pointed the current weaknesses of the two feature selection methods and put forward corresponding improving measures.

Keywords: Page categorization; Feature selection; Mutual Information algorithm; Chi-square algorithm

随着信息技术时代的到来，信息通讯技术的迅速发展，图书馆界也紧跟时代潮流，从最传统的藏书阁逐步发展到无墙图书馆、虚拟图书馆。直至 20 世纪 80 年代末以美国为首的西方国家逐渐形成统一意见，数字图书馆的概念才被统一下来，随后快速的在全球范围内发展起来。简单的说，数字图书馆就是将馆藏资源数字化，然后将数字化资源重新分类、整理，依托互联网为更多用户提供文献资源服务。数字图书馆在数字化资源过程中，如何维持馆藏资源的原貌，如何保证文献资源分类的准确，文本分类在这个过程中就变得至关重要。

1 文本分类技术简介

所谓文本分类,是指对所给出的文本,给出预定义的一个或多个类别标号,对文本进行准确、高效的分类，它是许多数据管理任务的重要组成部分[1]。通俗地理解，文本分类就是在限定的分类体系中，根据文本的内容或特征，分别将其划分到一个或者多个类别中。

总的来说，文本分类由文本预处理、文本表示、分类方法及效果评估 3 部分组成，主要涉及向量空间模型、特征提取和文本分类方法等关键技术。其中在进行文本分类之前，需要事先将文本的内容特征明显的表示出来，因此特征提取成为文本分类中的重要一环，能够直接影响文本分类的正确率。

2 数字图书馆文本分类中的特征提取方法

从上面的论述中可以看出想要进行文本分类，首要任务是找到一种能够减小向量空间维数又不影响文本分类效果的方法。如何从特征向量空间里剔除那些信息量少的词汇，已达到向量空间维数的减小，就是文本分类中的特征提取。

目前，常用的特征提取方法有特征频率（Term Frequency，TF）、文档频率（Document Frequency，DF）、信息增益（Information Gain，IG）、期望交叉熵（Expected Cross Entropy，ECE）、特征熵（Term Entropy）、互信息（Mutual Information，MI）和 χ^2 统计量（Chi-square，CHI）算法等。现重点分析互信息算法和 χ^2 统计量算法在数字图书馆文本分类中的应用，

收稿日期：2014-01-23

基金项目：菏泽学院人文社会科学研究项目“社交网络 SNS 在高校图书馆服务中的应用研究”（项目编号：XY12SK07）

作者简介：鲍凌云，菏泽学院图书馆助理馆员，硕士，研究方向：信息资源管理与科技创新。

指出传统算法中目前存在的不足并提出相应的改进措施，达到特征向量空间降维的目的，从而提高分类的查全率和查准率，进一步提高数字图书馆文本分类的准确率。

2.1 互信息算法（Mutual Information，MI）

互信息（Mutual Information）是信息论中的一个基本概念，用来度量两个随机变量间的统计依赖性或者一个变量包含另一个变量的信息量[2]。基于互信息的特征选取MI算法也是文本分类中常用的特征提取方法，MI算法使用公式(1)表示特征 t_k 与类别 C_i 的相关性：

$$MI(t_k,C_i)=\log\frac{P(t_k,C_i)}{P(t_k)P(C_i)} \quad (1)$$

在上面公式中，$P(t_k,C_i)$为类别 C_i、特征 t_k 同时出现的概率，$P(t_k)$为特征 t_k 出现的概率，$P(C_i)$为类别 C_i 出现的概率，$P(t_k|C_i)$为特征 t_k 在类别 C_i 中出现的概率。

从公式(1)可以看出，MI算法衡量的是某个词和类别之间的统计独立关系。对每个独立词，是以它在每个类别中的出现概率占它在整个文本集中的出现概率的比值作为它对每个类别的贡献[3]。

2.2 χ^2 统计量算法（Chi-square，CHI）

χ^2 统计量的概念来自列联表检验（Contingency Table Test），用来衡量特征 t_k 与类别 C_i 之间的统计相关性[4]。χ^2 统计量算法先假设特征 t_k 与类别 C_i 之间符合具有一维自由度的 χ^2 分布。然后利用特征 t_k 在类别 C_i 和其它类别 $\overline{C_i}$ 中出现的频率，来判断是否属于类别 C_i[3]。

特征 t_k 与类别之间的相关性定义如公式（2）：

$$CHI(t_k,C_i)=\frac{n[P(t_k,C_i)\times P(\overline{t_k},\overline{C_i})-P(\overline{t_k},C_i)\times P(t_k,\overline{C_i})]^2}{P(t_k)\times P(C_i)\times P(\overline{t_k})\times P(\overline{C_i})} \quad (2)$$

其中，n为文本数，$P(t_k,C_i)$为特征 t_k 出现的概率并且满足特征 t_k 属于类别 C_i，$P(\overline{t_k},\overline{C_i})$为特征 t_k 不出现的概率并且满足特征 t_k 不属于类别 C_i，$P(t_k,\overline{C_i})$为特征 t_k 出现的概率并且满足特征 t_k 不属于类别 C_i，$P(\overline{t_k},C_i)$特征 t_k 不出现的概率并且满足特征 t_k 属于类型 C_i。

从公式中可以看出，当 t_k 和 C_i 互相独立时，CHI值为0。可见CHI值能够直接反应特征 t_k 和类别 C_i 的相关程度，因此可以用 χ^2 统计量算法来衡量类别和特征的关联程度，从而帮助确定特征词的排除与保留。

3 数字图书馆文本分类特征提取方法的改进

3.1 互信息MI算法的改进

从MI算法的定义公式（1）可以看出，若计算出来 $P(t_k|C_i)$ 的值为绝对值很大的负值，如果按互信息算法统计排序，就容易把 $P(t_k|C_i)$ 为大负值的特征 t_k 排除，但是实际上，这样计算出来的结果只能说明特征 t_k 通常不出现在类别 C_i 中，但不能排除 t_k 在其它类别中的重要性，经过传统的MI算法计算，这样的特征 t_k 就容易被淘汰，显然这需要进一步改进，即保留MI值为绝对值大的负值的特征项。此外，MI算法的另一个问题是会出现特征词在类别中的概率不同，但利用公式（1）计算后得到相同的MI值，经过MI算法计算后，也容易在特征词的取舍上存在问题。根据现存问题，笔者对MI算法进行了逐步改进，详见表1。

由表1可见，根据文章指出的传统互信息算法的两点不足，分别通过取绝对值和添加概率系数的方法加以改进，改进后一来可以防止有用特征词的遗漏，二来也使得文档中特征间的类别区分更加明显。

3.2 χ^2 统计量算法的改进

根据 χ^2 统计量算法的公式，如果计算出的CHI值，一个非常接近0，而一个非常大，再根据CHI值排序，CHI值接近0的特征词很可能被除去，而CHI值非常大的特征词自然就被选中。这样容易导致的一种情况是，CHI值低的低频词往往因为所属文档的专有性，在被测文档中可能出现CHI值偏低的情况，但实际上是代表性很强；而CHI值高的高频词也有可能是在多类

表1 互信息算法的改进

原算法		$MI(t_k,C_i)=\log\frac{P(t_k\|C_i)}{P(t_k)}$
改进措施	目的：改善当MI值为负时，即使负值很大，特征词也被去掉	$MI(t_k,C_i)=\left\|\log\frac{P(t_k\|C_i)}{P(t_k)}\right\|$
	目的：防止概率不同的词在特定的文档中拥有相同的MI值	$MI(t_k,C_i)=P(t_k\|C_i)\log\frac{P(t_k\|C_i)}{P(t_k)}$
新算法		$MI^*(t_k,C_i)=P(t_k\|C_i)\left\|\log\frac{P(t_k\|C_i)}{P(t_k)}\right\|$

表2 X^2统计量算法的改进

原算法	$CHI(t_k, C_i)=\frac{n[P(t_k,C_i)\times P(\overline{t_k},\overline{C_i})-P(\overline{t_k},C_i)\times P(t_k,\overline{C_i})]^2}{P(t_k)\times P(C_i)\times P(\overline{t_k})\times P(\overline{C_i})}$	
改进措施	目的：降低在多类中普遍出现的高频词的权重，排除无用信息	$CHI'(t_k, C_i)=\log\frac{1}{P(t_k)}CHI$
	目的：提高对类别有意义的特征的排名，更准确地提取特征信息	$CHI^*(t_k, C_i)=P(t_k\mid C_i)\log\frac{1}{P(t_k)}CHI$
新算法	$CHI^*(t_k, C_i)=P(t_k\mid C_i)\log\frac{1}{P(t_k)}CHI$	

文档中普遍出现的词，相应的特征区分度就不高，代表性也不够强。如果按传统的χ^2统计量算法计算得到CHI值排序，就容易淘汰代表性很强的低频词，却保留代表性很弱的高频词，这样会严重影响文本分类的准确度。文章针对简单根据CHI值大小排序的χ^2，提出了改进措施，见表2。

经过表2中提出的两步改进，χ^2统计量算法可以说一定程度上克服了CHI值过大或过小对特征词淘汰与否的影响，能够更加有效地排除无用的特征信息，提取出对类别有意义的特征信息。

4 结束语

通过对互信息算法和χ^2统计量算法进行适度改进，一定程度上克服了传统算法中存在的不足。在数字图书馆中应用改进的互信息算法和χ^2统计量算法进行文本分类，理论上可以提高文本分类的准确率，为数字图书馆用户更方便、快捷的检索利用图书馆资源打好基础。

参考文献：

[1] 郭琛.数字图书馆中的中文网页文本分类器研究[D].武汉:武汉理工大学,2005.

[2] 冯学智等.遥感数字图像处理与应用[M].北京:商务印书馆,2011:234-235.

[3] 朱丽娜.中文网页分类特征提取方法研究[D].北京:中国石油大学,2009.

[4] 熊忠阳,张鹏招,张玉芳.基于χ^2统计的文本分类特征选择方法的研究[J].计算机应用,2008,(2):513-514.

数字图书馆文本分类特征提取方法研究与改进

作者： 鲍凌云， BAO Ling-yun
作者单位： 菏泽学院图书馆,山东菏泽,274015
刊名： 农业图书情报学刊
英文刊名： Journal of Library and Information Sciences in Agriculture
年，卷(期)： 2014,26(7)

参考文献(4条)

1. 郭琛 数字图书馆中的中文网页文本分类器研究 2005

2. 冯学智 遥感数字图像处理与应用 2011

3. 朱丽娜 中文网页分类特征提取方法研究 2009

4. 熊忠阳;张鹏招;张玉芳 基于x2统计的文本分类特征选择方法的研究 2008(02)