文章编号：1001－9081(2008)02－0513－02

基于 χ^2 统计的文本分类特征选择方法的研究

熊忠阳，张鹏招，张玉芳
(重庆大学 计算机学院，重庆 400044)
(zyxiong@cqu.edu.cn)

摘 要：特征提取是文本分类过程中的一个重要环节，它的好坏将直接影响文本分类的准确率。在研究文本分类特征提取方法的基础上，分析了 χ^2 统计的不足，并提出将频度、集中度、分散度应用到 χ^2 统计方法上，对 χ^2 统计进行改进，并通过实验对比改进前后的方法对文本分类效果的影响。在实验中，改进方法的分类效果要好于传统方法，从而验证了改进方法的有效性和可行性。

关键词：特征提取；χ^2 统计；频度；集中度；分散度
中图分类号：TP391　　文献标志码：A

Improved approach to CHI in feature extraction

XIONG Zhong-yang, ZHANG Peng-zhao, ZHANG Yu-fang
(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: Feature extraction technology is an essential part of text categorization, which directly affects the categorization precision. This paper comprehensively took frequency, distribution and concentration into account and proposed an improved Chi-square Statistic(CHI) approach. In order to verify the improved CHI approach, a contrastive experiment was carried out. The experimental results show that improved CHI approach is superior to traditional CHI approach in feature selection, which verifies the efficiency and probability of the improved CHI approach.

Key words: feature extraction; CHI approach; frequency; concentration; distribution

0 引言

文本分类面临的首要问题就是如何在计算机中合理的表示文本，这种表示方法既要包含足够的信息反映文本的特征，又不至于太过庞大使学习算法无法处理。最常用的文本特征表示模型是向量空间模型（Vector Space Model, VSM）。向量空间模型基于这样一个关键假设，即文章中词条出现的顺序是无关紧要的，它们对于文本的类别所起的作用是相互独立的，因此可以把文本看作一系列无序词条的集合。在该模型中，文本空间被视为一组正交词条向量所张成的向量空间[1]。向量的维数往往是惊人的，包含噪声，且特征不明显。特征提取可以看作是从测量空间到特征空间的一种映射或变换。特征提取可以降低特征空间的维数，从而达到降低计算复杂度和提高分类准确率的目的[6]。

对 VSM型的文本样本一般是构造一个特征评估函数，将测量空间的数据映射到特征空间，对特征空间中的特征值进行评估，然后选择合适的词作为样本的特征。特征评估函数通常有下列几种形式：文档频率（Document Frequency, DF）、信息增益（Information Gain, IG）、互信息（Mutual Information, MI）及词条的 χ^2 统计（CHI）等。

文献[5]指出 IG和 CHI的效果最好，IG计算量相对其他几种方法较大，因此本文主要针对特征选择的 χ^2 统计方法对其进行研究和改进。

1 χ^2 统计方法

χ^2 统计量的概念来自列联表检验（Contingency Table Test），它可以用来衡量特征 t 和类别 c 之间的统计相关性[3]。χ^2 统计方法度量词条 t 和文档类别 c 之间的相关程度，并假设 t 和 c 之间符合具有一阶自由度的 χ^2 分布。如果 A 表示包含词条 t 且属于类别 c 的文档频数，B 为包含 t 但是不属于类别 c 的文档频数，C 表示属于类别 c 但是不包含 t 的文档频数，N 表示语料中文档总数，D 表示既不属于 c 也不包含 t 的文档频数。则对于 t 的 χ^2 值由式(1)[3]计算：

$$\chi^2(t,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

考虑到 N, $A+C$, $B+D$ 均是常数，式(1)可以简化为：

$$\chi^2(t,c)=\frac{(AD-BC)^2}{(A+B)(C+D)} \tag{2}$$

当特征 t 与类别 c 相互独立时，$\chi^2(t,c)=0$，此时特征 t 不包含任何与类别 c 有关的鉴别信息。特征 t 与类别 c 的统计相关性越强，$\chi^2(t,c)$ 的值就越大，此时特征 t 包含的与类别 c 有关的鉴别信息就越多。

对于多类问题，分别计算 t 对于每个类别的 χ^2 值，再用式(3)计算词条 t 对于整个语料的 χ^2 值，分别进行检验：

$$\chi^2_{MAX}(t)=\max_{1\le i\le m}\{\chi^2(t,c_i)\} \tag{3}$$

其中 m 为类别数。从原始特征空间中移除低于特定阈值的词条，保留高于该阈值的词条作为文档表示的特征。本文就是采用式(3)计算词条 t 对于整个语料的 χ^2 值。

由 χ^2 的计算公式可以看出，χ^2 统计方法只考虑了特征在所有文档出现的文档频数，并没有考虑特征在某一文档中出现的频率。如果某一特征只在一类文档的少量文档中频繁出现，通过 χ^2 计算公式计算出来的 χ^2 统计值很低，在特征选择

收稿日期：2007－09－10；修回日期：2007－11－10。　基金项目：重庆市科委自然科学基金资助项目（CSTC2006BB2021）。
作者简介：熊忠阳（1964－），男，重庆人，教授，博士生导师，博士，主要研究方向：数据挖掘、数据库、并行计算、网络信息处理；张鹏招（1983－），男，湖南浏阳人，硕士研究生，主要研究方向：文本分类、互联网应用技术研究；张玉芳（1967－），女，上海人，副教授，博士，主要研究方向：数据挖掘、数据库、并行计算、网络信息处理。

时这种特征词就会被排除掉, 但是这种在少量文档中频繁出现的特征词很有可能对分类的贡献很大, 比如专指概念。这是 χ^2 统计不足之一, 它对低文档频的特征项不可靠。

统计值基于 χ^2 分布, 特征词的 χ^2 统计值比较了特征项对一个类别的贡献和对其他类别的贡献[7], 这样就可能把对其他类别贡献大特征项选择出来。这种特征项往往是在指定类中出现频率较低而普遍存在于其他类, 它们对分类的贡献不大, 应该删除。这也可以由 χ^2 的计算公式看出, 当 BC> AD, 也就是特征词在其他类别频繁出现而在指定类中很少存在时, 计算出来的 χ^2 统计值很高。这也是 χ^2 统计不足之处, 它提高了在指定类中出现频率较低而普遍存在于其他类的特征项在该类中的权重。

2 χ^2 统计方法的改进

特征抽取算法的优劣直接影响到文本分类的效果, 特征项选择依赖于频度、分散度和集中度等多项测试指标[8]。通过前文对 χ^2 统计方法不足的分析, 我们总结出特征提取的关键在于提取出集中分布在某类文档中而且在这类文档中均匀分布频繁出现的特征, 于是本文综合考虑频度、集中度、分散度[2]等三项测试指标对 χ^2 统计方法进行改进。

2.1 频度

频度是最常用的指标, 采用该指标是基于这样的想法: 在某一类文本中出现次数越多的特征项越能代表这类文本, 因此选择在同一类文本中出现频度最高的若干特征项作为该类文本的类别特征。设训练集中类别为 C_j 的文本有 $d_1, \cdots d_k \cdots, d_m$, 特征 t 在文本 $d_k(1 \leqslant k \leqslant m)$ 中出现的频度为 tf_k 则特征项 t 在类别 C_j 中出现的频度 α 表示如下:

$$\alpha = \sqrt{\sum_{k=1}^{m} (tf_k)^2} \tag{4}$$

引入频度来改进 χ^2 统计方法主要是为了解决 χ^2 统计方法对低文档频特征项不可靠的问题。

2.2 集中度

采用集中度指标是基于这样的想法: 一个有标引价值的特征项, 应该集中出现在某一类文本中, 而不是均匀地分布在所有各类文本中。设类 C_j 中包含特征项 t 的文档个数为 df_j 则特征项 t 在类别 C_j 中出现的集中度 β 表示如下:

$$\beta = \frac{(df_j - tf_j)^2}{tf_j}, \quad tf_j = \frac{\sum_{i=1}^{n} df_j}{n} \tag{5}$$

其中 n 是文本类别总数。

2.3 分散度

采用分散度指标是基于这样的想法: 在某类文本中均匀出现的特征项对该类文本应具有较高的标引价值, 若只集中出现在该类的个别文本中, 而在该类别的其他文本中很少出现, 则该词的标引价值就要小多。设类 C_j 中包含特征项 t 的文档个数为 df_j 则其分散度表示如下:

$$\gamma = df_j \tag{6}$$

分散度用特征项在某一类中出现的文档频数表示, 分散度越大表明特征项在某一类的文档中出现次数越多, 则该特征项在这一类中分布越均匀。

由以上定义可知, 对于某一特征项, 其频度越高、集中度越强、分散度越大, 则对文本分类越有用, 即分辨度越强。于是本文在 χ^2 统计方法的基础上加上频度、集中度、分散度的修正, 对 χ^2 统计方法作出改进, 得到如下的改进公式:

$$\chi^2(t, c) = \frac{(AD - BC)^2}{(A+B)(C+D)} \times \frac{\alpha + \beta + \gamma}{3} \tag{7}$$

改进方法引入频度测试指标使得在计算特征项的 χ^2 统计值时考虑了特征项在文档内的出现频率, 解决了传统的 χ^2 统计方法对低文档频特征项不可靠的问题; 引入集中度和分散度测试指标使得选择出来的特征项尽可能是集中分布在某类文档中而且在这类文档中均匀分布的特征项, 解决了传统的 χ^2 统计方法提高了在指定类中出现频率较低而普遍存在于其他类的特征项在该类中的权重的问题。

3 实验与分析

在研究文本分类的过程中, 特征提取是最关键的环节之一, 具有降低向量空间维数、简化计算、防止过分拟合以及去除噪声等作用, 特征提取的好坏将直接影响文本分类的准确率。

3.1 文本分类

文本分类是信息检索领域中一个极为重要的子问题, 特别是在可以作为半结构化网上文档信息极为丰富的情况下, 以文本分类为基础的各种应用, 如个人信息代理、搜索引擎、网上信息发布等, 已成为有效控制和利用海量信息的重要手段。现今已有诸多新分类技术和方法被提出来, 例如 Expert Networks 支持向量机的文本分类, 神经网络的文本分类, Windrow-Hoff和 EG等[4]。

由于本文的研究对象是特征提取方法的改进, 所以在此对各种分类方法不做详细讨论, 仅选取最为常用的朴素贝叶斯方法来进行分类。Naive Bayes分类方法是一种简单而又非常有效的分类方法。该算法的基本思路是计算文本属于类别的概率, 文本属于类别的几率等于文本中每个词属于类别的几率的综合表达式。计算公式如下:

$$P(w_i \mid c_j) = \frac{w_i \text{在 } c_j \text{类别文档中出现的次数}}{\text{在 } c_j \text{类所有文档中出现的次数}} \tag{8}$$

3.2 实验及结果

本文从来源于搜狐新闻网站的大量经过编辑手工整理与分类的新闻语料中抽取部分短文共计 2700篇, 采用交叉验证的方法对 9个类的文档进行实验。

表 1 文本分类中传统的与改进的 χ^2 统计方法实验结果比较表

χ^2 统计	参数	体育	艺术	军事	财经	教育	健康	招聘	计算机	旅游
传统的方法	R	0.870	0.813	0.883	0.760	0.587	0.763	0.643	0.733	0.763
	P	0.891	0.701	0.904	0.838	0.739	0.815	0.720	0.530	0.784
改进的方法	R	0.930	0.867	0.920	0.770	0.703	0.807	0.890	0.820	0.780
	P	0.924	0.734	0.936	0.920	0.868	0.893	0.893	0.594	0.863

最终的实验结果依靠通用的召回率 (recall) 和正确率 (Precision) 两个指标加以衡量和比较。传统的与改进的 χ^2 统计方法实验结果如表 1所示。

由实验结果比较分析, 改进的 χ^2 统计方法进行分类时的召回率和正确率都较传统的 χ^2 统计方法好。

(下转第 518页)

4 应用研究

在军工制造业中，主机制造厂为了在有限时间内生产大量武器装备，来协同其他具有特定制造能力的制造单元来共同完成任务，因而扩散制造的运作模式就在目前军工制造业中应运而生。扩散制造将武器装备产品划分成若干模块、次模块、组件、附件和零件等部分，对各部分的制造工艺进行优化与固化，然后把生产任务扩散到具有特定制造能力的制造单元，充分利用资源的整合与优化所产生的集群效应，实现武器装备在较短时间内完成大批量的生产任务。扩散制造的生产模式要求企业在网络化制造环境下能够快速响应市场需求，需要决策者能够快速制定相应的决策。我们在某扩散制造企业的决策支持系统中利用一 B/S模式的分布式 Web-OLAP系统。该扩散制造决策支持系统中某个扩散件一次交货合格率同期对比，如图 4所示。

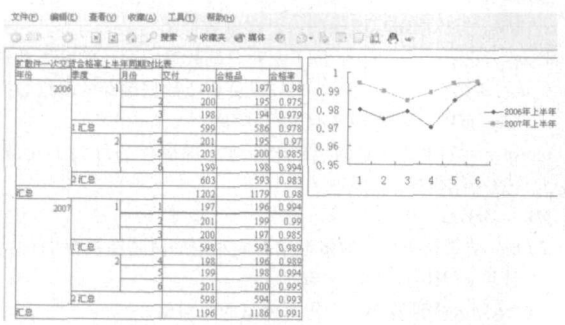

图 4 扩散件一次交货合格率同期对比

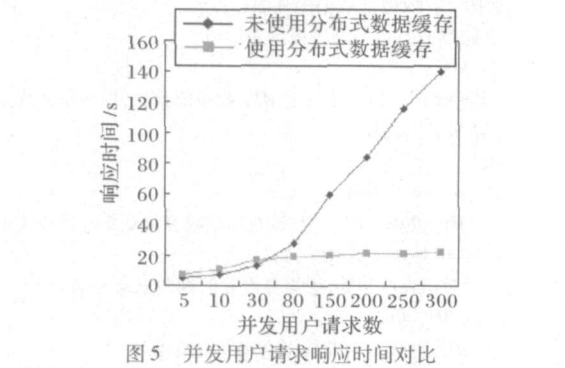

图 5 并发用户请求响应时间对比

将一般情况下的决策支持系统与本文的操作结果进行对比后如图 5所示，本文所采用的方法可以在用户数量增多时，较高地提高系统的响应速度。

5 结语

本文将分布式数据缓存技术引入到 Web-OLAP系统中，通过总部数据缓存，与各分部的数据缓存协同工作，在客户端用户数据增多的情况下，减轻各服务器的查询负担，充分利用网络中的资源，缩短客户在做查询分析时的响应时间。

参考文献：

[1] 王珊. 数据仓库技术与联机分析处理[M]. 北京：科学出版社，1998.

[2] 王惠敏. 网络环境下对分布式决策支持系统的探讨[J]. 价值工程，2006，8：88－90.

[3] MADEIRA H，COSTA J，VIEIRA M. The OLAP and data warehousing approaches for analysis and sharing of results from dependability evaluation experiments[C]// International Conference on Dependable Systems and Networks. [S.l.]：IEEE Press，2003：86－91.

[4] YANG W，ZHU W，LIU Y. Research of a Web-based DSS intelligent Agent over data warehouse[C]// 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI'04). Beijing：IEEE Press，2004：433－436.

[5] POWER D，KAPARTHI S. Building Web-based decision support systems[J]. Studies in informatics and control，2002，11(4)：291－302.

[6] 于雅丽，谢强，丁秋林. Web环境下基于对象池和数据缓存技术的 OLAP系统[J]. 武汉大学学报：工学版，2006，39(6)：59－62.

[7] 赵玉伟. WWW中缓存机制的应用研究[D]. 武汉：武汉理工大学，2006.

（上接第 514页）

同时应该指出，尽管改进的 χ^2 统计方法取得了一定效果，但针对整个文本分类问题的效果仍未出现明显提高。正如文献[5]所讲：文本分类问题是涉及到文本表示、相似度计算和算法决策等多种复杂技术的综合应用。

4 结语

为了解决 χ^2 统计方法提高了在指定类中出现频率较低，却又普遍存在于其他类的特征项在该类中的权重以及对低文档频的特征词不可靠两个不足之处，本文综合考虑了频度、集中度、分散度等三项指标，提出了改进的 χ^2 统计方法。经过文本分类系统的实验验证，改进的 χ^2 统计方法性能要好于传统的方法。

参考文献：

[1] 李凡，鲁明羽，陆玉昌. 关于文本特征抽取新方法的研究[J]. 清华大学学报：自然科学版，2001，41(7)：98－101.

[2] 陈治纲，何丕廉，孙越恒，等. 基于向量空间模型的文本分类系统的研究与实现[J]. 中文信息学报，2004，19(1)：36－41.

[3] SCHUTZE H，HULL D A，PEDERSEN J O. A comparison of classifiers and document representations for the routing problem[C]// Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval. [S.l.]：ACM Press，1995：229－237.

[4] 鲁松，李晓黎，白硕，等. 文档中词语权重计算方法的改进[J]. 中文信息学报，2000，14(6)：8－13.

[5] YANG Y，LIU X. Expert network：Effective and efficient learning from human decisions in text categorization and retrieval[C]// Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin：Springer，1994：12－22.

[6] 徐凤亚，罗振声. 文本自动分类中特征权重算法的改进研究[J]. 计算机工程与应用，2005，41(1)：181－184.

[7] 程泽凯，陆小艺. 文本分类中的特征选择方法[J]. 安徽工业大学学报，2004，21(3)：221－224.

[8] 杨允信. 文本文件自动分类之研究[C]//台湾地区第六届计算语言学研讨会论文集. 台湾：[s.n.]，1993.