第34卷 第23期 计 算 机 工 程 2008年12月
Vol.34 No.23 Computer Engineering December 2008
·软件技术与数据库· 文章编号：1000—3428(2008)23—0089—03 文献标识码：A 中图分类号：TP311.5

中文文本体裁分类中特征选择的研究

邓 琦，苏一丹，曹 波，闭剑婷
(广西大学计算机与电子信息学院，南宁 530004)

摘 要：针对文本体裁自动分类在特征选择和权重计算方面的特殊性，提出文本的内容类别信息，改进传统特征选择方法CHI以及权重计算公式tf.idf，并运用支持向量机在含5类体裁的语料上进行中文文本体裁自动分类。实验结果表明，该方案是可行的。
关键词：中文信息处理；体裁分类；特征项选择；支持向量机

Research on Feature Selection in Chinese Text Genre Classification

DENG Qi, SU Yi-dan, CAO Bo, BI Jian-ting
(College of Computer and Electronic Information, Guangxi University, Nanning 530004)

【Abstract】 Aiming at the particularity of text genre classification in feature selection and weight calculation, this paper presents the text content category information, which improves the conventional CHI feature selection method and the tf.idf formula of feature weight. By using Support Vector Machine(SVM), an automatic classification on a Chinese text corpus consisting of five genres is carried out. Experimental results show this scheme is feasible.
【Key words】 Chinese information processing; genre classification; feature selection; Support Vector Machine(SVM)

1 概述

随着万维网内信息量的增加以及人们对个性化搜索需求的增大，文本体裁自动分类问题已成为当前计算语言学及传统语言学研究的热点之一。

文本体裁自动分类就是由机器从文章中分析出文本体裁的过程，它与把主题表达出来形成文章的写作过程互逆。写作一般分为确定主题、构思和形成文本3个步骤，其中，构思的任务就是要选择合适的体裁和形式来表现内容。文本体裁既受到主题的制约，又对词汇起着规范和先导作用，所以，体裁分类与文本中的主题信息、词汇信息都密切相关。

目前，较成熟的主题分类仅考虑词汇信息，本文借鉴主题分类的研究成果探索体裁分类，并针对体裁分类的特殊性引入文本的内容信息，并进行中文文本体裁自动分类的研究实验。

2 体裁分类研究概述

文本自动分类是将训练文本表示成计算机可识别处理的结构，再由计算机建立文本属性和文本类别之间的关系模型，即可利用该模型判断新文本类别。向量空间模型是目前被证实效果较好的文本表示模型之一。该模型将文本d_i表示成n维向量$V(d_i)=(t_{i1},w(t_{i1}); t_{i2},w(t_{i2});\cdots; t_{in},w(t_{in}))$，其中，$t_{i1}, t_{i2},\cdots, t_{in}$为特征；$w(t_{ik})$, $k=1,2,\cdots,n$为第k个特征的权重。体裁分类与主题分类在分类模型和算法方面无本质区别，两者的区别主要集中在特征选择和权重计算方面[1]。

和仅使用字、词为特征的主题分类不同，体裁分类还需使用能体现文章结构与风格的复杂特征。Kessler[2]将抽取体裁特征的线索分成句法结构线索、词汇线索、字符级线索和派生线索，依次选定55个特征项并在规模为402个手工分类的训练文本语料库上取得较好的分类效果。其后很多实验都是在这4类线索的基础上展开特征项选择研究[3-4]。

特征选择还需用评价函数为各候选特征项打分，按分值大小对候选项进行排序并选取预定数目的特征项以达到降维目的。常用的评价方法包括：χ^2统计量法(CHI)和相关系数法。

令A表示训练集中项t和类别c同时出现的次数；B表示t出现c不出现的次数；C为t不出现c出现的次数；D为t和c均未出现的次数；n为训练集中的样本总数，则t对于c的χ^2统计量公式为

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

χ^2统计值越高，t与c的独立性越小，相关性则越强。

相关系数法认为特征项的重要性应由其正相关能力决定，故对χ^2统计量公式进行开方运算并保留正负号，以区别特征项对类别的贡献是正贡献还是负贡献，同时缩小负贡献的影响，其表达式为

$$CC(t,c)=\frac{\sqrt{n}(AD-CB)}{\sqrt{(A+C)(B+D)(A+B)(C+D)}}$$

为特征项赋权是为了衡量不同特征项对文档的重要程度和区分度，常用的赋权函数tf.idf公式为

$$w(t_i,d_j)=tf(t_i,d_j)\times \mathrm{lb}\left(\frac{N}{n_i}+0.01\right)$$

其中，词语频率(term frequency, tf)为词t_i在文档d_j中出现的频率；词语倒排文档频率(inverse document frequency, idf)为词t_i在文档集中的分布频率；N为文档集中的文档数；n_i为含词t_i的文档数；idf值越小说明该词越普遍，赋予的总权值也越小。

3 中文文本体裁自动分类中的特征选择研究

传统特征选择方法和权重公式都是围绕特征项与类别间的关系展开研究的，这些方法在主题分类中取得了较好的分

基金项目：国家自然科学基金资助项目(60564001)
作者简介：邓 琦(1982－)，女，硕士研究生，主研方向：自然语言处理；苏一丹，教授；曹 波、闭剑婷，硕士研究生
收稿日期：2008-04-11 **E-mail：**dengqi_cn@yahoo.com.cn

类效果。体裁分类与主题分类的差别在于其特征选择和权重计算需要同时考察词汇、体裁类别以及内容类别三者之间的关系。

3.1 改进的χ^2统计量法$\chi^2(t,c,g)$

传统χ^2统计量法和相关系数法都是卡方公式在独立性检验方面的应用，主题分类常用它们来考察词汇t和内容类别c之间的统计相关性强度，但直接套用这些公式会使词汇脱离内容类别信息的约束。若能在体裁分类的特征选择中引入内容类别信息，即先将词汇与内容类别信息联系在一起，再考察其与体裁类别之间的关联性，选取出的词汇特征项将更具体裁辨识能力，更符合人工经验。

为将内容类别信息合理引入体裁分类中，先将内容类别对体裁类别的影响程度相对量化。较直观的方法就是观察各内容类在所有体裁类中的文本分布曲线差异情况：在不同体裁类别中分布差异越明显的内容类，区分体裁的贡献越大。这里随机抽取 250, 500, 750 和 1 000 篇语料，观察汽车、教育、生活、旅游、科技、娱乐、军事、财经、房产、体育等十余种内容类在各体裁类上的分布来考察 2 种类别间的关系。除去出现次数较少的游戏类，随语料规模变化而分布变化较大的军事、教育类以及在各体裁中分布差异不明显的财经、科技、汽车类，得到在体裁类别中分布差异较大，可认为对体裁分类有较优辨别作用的可标识内容类别旅游、生活、娱乐、体育和房产。

由于受主题分类准确度的影响，因此对内容类别与体裁类别间的约束只得到粗粒度量化，根据实验结果可将内容类粗略地划分成可标识内容和非可标识内容 2 种类别。可标识内容类别对体裁分类辨别能力大于非可标识内容类别，如娱乐类的文本属于新闻体裁的概率大于应用文，同样，房产也极少出现在应用文中。图 1 是在 500 篇语料的实验中，各可标识内容类别在各体裁类别中的文档分布情况。

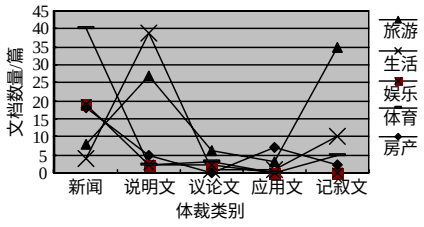

图 1　内容类文档在体裁类别中的分布

为考察词汇、内容类别以及体裁类别三者的联系而又不引入维数灾难，本文先将词汇与内容信息简单加合，其中，内容信息根据以上分析结果分为可标识内容和非可标识内容 2 类，再根据卡方公式考察复合事件 X 与体裁类别间的关联程度。设变量 $X=\{H1, H2\}$, $Y=\{T1, T2\}$。

事件 $H1$ 为词汇 t 与可标识内容类别 c 相关联。为简化问题，将词汇 t 对可标识内容类别 c 的正相关能力与负相关能力同等对待，即文本中 t 与 c 同时出现和同时不出现的情况都视为两者关联。事件 $H2$ 为词汇 t 与可标识内容类别 c 不相关联。$T1$ 为文本属于体裁类别 g。$T2$ 为文本不属于 g。显然有 $H1+H2=T1+T2=N$, N 为文本总数。

根据卡方值计算公式，得到以下公式：

$$\chi^2=\frac{N\left[(n_1+n_5)(n_6+n_7)-(n_4+n_8)(n_2+n_3)\right]^2}{(n_1+n_2+n_3+n_5)(n_1+n_4+n_5+n_8)(n_2+n_3+n_6+n_7)(n_4+n_8+n_6+n_7)}$$

其中，$n_i(i=0,1,\cdots,7)$表特征t、可标识内容类别c及体裁类别g的 8 种共现情况的出现次数。8 种共现情况依次为(t,c,g); $(\_t,c,g)$; $(t,\_c,g)$; $(t,c,\_g)$; $(\_t,\_c,g)$; $(t,\_c,\_g)$; $(_t,c,_g)$和$(_t,_c,_g)$, $N=n_1+n_2+n_3+n_4+n_5+n_6+n_7+n_8$为文本总数。通过检测同一文本中 2 个变量$X$与$Y$的关联程度，得到词汇、可标识内容类别与体裁类别三者的关联程度，理论上χ^2值越大，三者关联程度越高，赋予该词汇的评估分值也越高。

3.2 改进的 tf.idf 权重公式 tf.idf.icf

传统 tf.idf 权重公式只考虑 tf 和 idf 2 个因素，体裁分类中该公式不能准确反映词汇对体裁类别的区分程度，要将内容类别信息也引入权重公式中。为权重公式引入新的权重因素(inverse content frequency, icf)是基于这样一种假设：词语当前所隶属的内容类别在所有体裁类别中出现频率越小，词语对区别文档越有意义。改进的权重公式为

$$w(t_i,d_j)=tf.idf.icf$$

其中，$icf=\mathrm{lb}(\frac{M}{m_i}+2)$；$tf$和$idf$的意义与传统$tf.idf$权重公式相同；$M$为文档集中出现的体裁类别数；$m_i$为对于词语当前所隶属的内容类别$c_i$，文档集中出现过该$c_i$的体裁类别数。为过滤干扰数据，设当且仅当$c_i$出现在体裁类$g_j$中的数量大于等于$c_i$出现总数量的 2%时，才可记为“$c_i$在$g_j$中出现过”。考虑到该权重因素虽然对整个词汇的权重有影响，但影响程度有限，故对M/m_i加上一个较大值 2 再取对数，以减缓icf随自变量变化的剧烈程度及其对整个权重公式的影响程度。

icf 权重因素将词语与其隶属的内容类别联合起来考察，相当于把词语放入具体的内容语境中分析，引入该因素使权重公式更适合于体裁分类。例如，“政府”这个词常出现在科技和体育 2 不同内容的文本中，由于科技类在各体裁类别中分布都较平均，体育类则主要出现于新闻体裁中，易见属于体育类文本的情况下，“政府”这个词应更具有体裁类别的辨识能力，icf 权重的引入符合该人工经验。

4　实验与评测

4.1　语料库的构建

实验在哈工大 IR 研究室提供的单文档自动文摘中的应用文、新闻、记叙文、说明文、议论文 5 类体裁、211 篇语料的基础上，从互联网上补充 1 289 篇文章，并使用哈工大 IR 研究室提供的语言技术平台 LTP 接口进行命名实体识别、主题分类、浅层语义标注等预处理，与已有语料一起形成含 5 个体裁类，各 300 篇语料的中文文本体裁语料库。

4.2　特征项选择

实验所用特征项含 4 种共 148 项。(1)字符级特征：体裁会制约文本的句型，如应用文多用陈述句，少用疑问句。依此实验选用感叹号、疑问号等标点的出现次数及分布密度为特征。(2)浅层分析级特征：体裁对文本词汇亦有一定约束。如新闻常出现时间、人名、地名等专有名词；政府报告、开幕词等应用文常使用简称，大量出现机构名。根据这些线索提取的特征项包括文章长度、句子数量、平均句长、段落数量、平均段长、词汇数量及数词、拟声词、时间词、方位词、人名、机构名的频次等。(3)深层分析级特征：在LTP对文本内容进行自动分类的基础上，将内容类别以及是否为可标识内容类别等深层分析结果作为特征项引入实验。(4)词汇级特征：浅层分析级线索只将某类词汇作为特征项，脱离了词汇的词义信息。借鉴主题分类的思想取高频词汇为特征项。不同的是实验将词汇的选取缩小在具独立词义的名词范围内，使用 100 个高频名词为特征项。但对训练集进行分词预处理

后，得到的是上万个特征项，因此，还需用选择方法来对特征空间进行降维。实验分别使用传统χ^2统计量法($\chi^2(t,g)$)、相关系数法以及改进χ^2统计量法($\chi^2(t,c,g)$)进行特征χ选择，权重亦分别使用传统tf.idf和改进后的tf.idf.icf公式进行计算。

4.3 SVM 分类器

SVM 是近年使用较广且被证实在高维、小样本的训练空间中有较好分类性能的分类器。其基本思想是通过定义适当的内积将输入空间变换到高维空间，然后求解最优分类面并利用该分类面的分割特性对新文本进行分类。实验使用的分类器软件是由台湾大学林智仁副教授开发设计的 libsvm-2.81。

4.4 实验结果与分析

实验将语料按 1:1 随机分成训练集和测试集进行封闭测试和开放测试。分类结果用准确率和综合了准确率及召回率 2 指标的F_1测试值衡量。结果显示分类总体效果较好。封闭测试各指标均接近 100%。表 1、表 2 是开放测试中使用不同选择方法和权重公式得到的准确率P和测试值F_1。

表 1　使用 tf.idf 情况的分类结果　(%)

方法类	χ^2统计量法		相关系数法		改进的χ^2统计量法	
	P	F_1	P	F_1	P	F_1
应用文	83.46	87.02	82.05	83.33	86.45	89.01
新闻	87.50	89.16	87.23	87.78	88.10	88.23
记叙文	93.00	90.71	90.69	90.60	94.75	91.43
说明文	80.53	83.34	80.77	82.25	84.61	87.38
议论文	82.75	83.67	84.45	82.38	82.25	82.46
平均值	85.45	86.78	85.04	85.27	87.23	87.70

表 2　使用 tf.idf.icf 情况的分类结果　(%)

方法类	χ^2统计量法		相关系数法		改进的χ^2统计量法	
	P	F_1	P	F_1	P	F_1
应用文	91.01	91.94	89.37	87.04	91.48	90.52
新闻	91.92	92.82	93.87	92.00	91.02	89.36
记叙文	95.10	94.96	94.76	92.18	95.54	93.21
说明文	87.49	90.67	85.42	86.32	86.79	88.89
议论文	85.74	87.3	84.91	84.11	85.14	88.63
平均值	90.25	91.54	89.67	88.33	90.00	90.12

从表 1 中可以看出，同是使用传统权重公式的情况下，改进的χ^2统计量法的分类效果最好。对比表 1 和表 2 数据可以看出，在使用同种特征选择函数的情况下，改进的权重函数tf.idf.icf能使分类结果得到显著提高。但对改进的χ^2统计量法，搭配了改进的权重函数tf.idf.icf后的分类效果的提高程度比其他 2 种特征选择函数小，总效果反而弱于传统χ^2统计量法，但仍优于相关系数法。由于改进的χ^2统计量法受主题分类精度及标志与非标志性内容类别的界定等因素的影响，因此要体现其优越性还需提高这些因素的精确度。

5　结束语

文本体裁分类已在搜索引擎、数字化图书馆等诸多领域受到重视，主题分类的发展也为体裁分类研究提供了丰富的理论基础和分析工具。本文从字符级、词汇级、浅层分析级与深层分析级线索入手选定特征项，并引入内容类别信息对传统评估函数和权重函数进行改进。实验证实了中文文本体裁自动分类的可行性。由于特征选择和权重计算是体裁分类中的核心问题，因此如何抽取更复杂的深层分析级线索是今后体裁分类研究中要逐步解决的问题。

参考文献

[1] Chul S, Kong Joo Lee. Multiple Sets of Features for Automatic Genre Classification of Web Documents[J]. Information Processing & Management, 2005, 41(5): 1263-1276.

[2] Brett K, Geoffrey N, Hinrich S. Automatic Detection of Text Genre[C]//Proc. of the 35th Annual Meeting on Association for Computational Linguistics. Madrid, Spain: [s. n.], 1997.

[3] Yong Bae Lee, Hyon M. Text Genre Classification with Genre-revealing and Subject-revealing Features[C]//Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tampere, Finland: [s. n.], 2002.

[4] Aidan F, Nicholas K. Learning to Classify Documents According to Genre[J]. Journal of the American Society for Information Science and Technology, 2006, 57(11): 1506-1518.

（上接第 85 页）

正结果。目前还没有研究和工具能用从 JVMPI 获得的可观察信息推导出对象的行为及其使用逻辑模型，更不用说直接利用对象行为和使用逻辑来解决 Java 内存低效使用问题了。笔者认为这些研究和工具未能从 JVMPI 获得原始信息构造对象行为及对象使用逻辑信息，并利用它们检测和修正内存低效问题是导致这些研究和工具不足的主要原因。

6　结束语

目前，笔者在软件故障定位研究及分析Java程序内存低效使用的过程中提出了对象生存期行为模型(Lifetime Behavior Model, LBM)，并取得一定进展，同时验证了它的有效性[6]。基于LBM实现了表示对象活动状况的活动谱图模型(Activity Spectrum Model, ASM)，该模型能支持程序中对象内存使用性能的分析。下一步工作将研究基于对象行为模式的软件故障定位。

参考文献

[1] QUEST 公司. JProbe Memory Debugger Developer's Guide[Z]. 2005.

[2] Wim P, Sevitsky G. Visualizing Reference Patterns for Solving Memory Leaks in Java[C]//Proc. of the 13th European Conf. on Object-oriented Programming. Lisbon, Portugal: [s. n.], 1999.

[3] Shaham R, Kolodner E, Sagiv M. Automatic Removal of Array Memory Leaks in Java[C]//Proc. of the 9th International Conference. Berlin, Germany: [s. n.], 2000.

[4] Mitchell N, Sevitsky G. Leakbot: An Automated and Lightweight Tool for Diagnosing Memory Leaks in Large Java Applications[C]//Proc. of European Conf. on Object-Oriented Programming. Darmstadt, Germany: [s. n.], 2003.

[5] Larsen S. Memory Leaks, Be gone[Z]. (2005-06-27). http://dev2dev.bea.com/pub/a/2005/06/memory_leaks.html.

[6] Wu Ji, Jia Xiaoxia, Li Guohuan, et al. Java Object Behavior Modeling and Visualization[C]//Proc. of International Conf. on Software Engineering Advances. Tahiti, French: [s. n.], 2006.