维普资讯 http://www.cqvip.com

第34卷 第18期 计 算 机 工 程 2008年9月
Vol.34 No.18 Computer Engineering September 2008
·软件技术与数据库· 文章编号：1000—3428(2008)18—0087—02 文献标识码：A 中图分类号：TP309

实时文本分类系统的研究与实现

黄 旭，朱艳琴，罗喜召
(苏州大学计算机科学与技术学院，苏州 215006)

摘 要：分析文本分类过程中影响实时性的因素，即分词处理高耗时和特征空间维数过高问题。结合网页过滤的实时应用提出一种实时文本分类方法，弱化分词处理过程，降低特征空间维数，以提高分类速度。通过优化特征项选取维持分类效果，基于贝叶斯理论实现实时文本分类系统。实验结果表明，该方法在维持精确率和召回率分别为85%, 94%的情况下，显著提高了分类速度。

关键词：信息安全；内容安全；文本分类

Research and Implementation of Real-time Text Categorization System

HUANG Xu, ZHU Yan-qin, LUO Xi-zhao
(School of Computer Science and Technology, Soochow University, Suzhou 215006)

【Abstract】 This paper analyzes the factors which affect the quality of real-time in text categorization, that is the high time-consuming problem of word segmentation, and the excessively high dimension of character space. Based on the real-time application of Web filter, a real-time text categorization approach is proposed. The approach improves the rate of text categorization by reducing the processing of word segmentation and the dimension of character space. It maintains the effect of text categorization by optimizing the selection of character item, and implements a real-time text classifier based on Bayesian theory. Experimental results show that this approach improves the rate of text categorization effectively, and the precision and recall is maintained at 85 percent and 94 percent.

【Key words】 information security; content security; text categorization

1 概述

计算机和通信技术的快速发展使互联网成为重要的通信方式。随着互联网上信息量的不断增长，网络内容安全问题越来越严重，如何从 Internet 上迅速地提取反动、色情、暴力等不良信息，从而对网页实施有效监控具有重要现实意义。网页的自动分类是其中的关键技术之一，目前大多数网络信息表现为文本形式，因此，文本自动分类技术是网页自动分类的基础。

文本自动分类是一种两类分类技术，能将未知类别的文档划分到涉及国家安全敏感信息的文本集或常规文本集中。此类应用要求文本分类系统具有良好的分类效果及较高的处理速度。本文分析文本分类过程中影响实时性的因素，在保证分类精度的基础上，针对这些因素进行改进，设计并实现了一种可用于实时场合的两类文本分类系统。

2 实时文本分类方法

2.1 文本分类方法研究现状

文本分类系统的任务是在给定的分类体系下，根据文本内容自动确定文本关联的类别[1]，即为待分类文档设定类别标记。可用数学公式表示为

$$f: A \to B$$

其中，A 为待分类的文本集合；B 为分类体系中的类别集合。文本分类器一般由训练过程和分类过程2个逻辑过程组成[2]。训练过程通过文档实例集获取目标文档的特征表示，供分类算法使用。分类过程是将待分类文档以适当的数学模型表示，供分类算法识别。分类算法处于核心地位，而特征提取及其数学表示是文本分类的基础。文本分类器主要包括预处理、分词、特征提取、特征项数学表示、文本的数学表示、分类算法等模块。

目前文本分类方法分为2类，即基于规则的方法和基于概率统计的方法[3]。基于规则的方法归纳出训练样本中规律性的内容以形成规则，并根据此规则确定文本类别。此方法采用的主要算法有决策树、粗糙集、Ripper 方法、boosting 方法等。基于规则的分类方法在规律不明显的领域中应用效果较差。基于概率统计的方法统计出文档中用词等方面的概率分布规律，其本质也是获取一种分类规则，但这种规则不易被人理解。此方法采用的主要算法有 K 邻近方法、贝叶斯方法、Rocchio 方法、支持向量机等。

2.2 贝叶斯方法概述

贝叶斯方法被广泛应用于需要具备学习能力的智能系统中，其理论基于如下假定：待考查的量遵循某概率分布，根据这些概率及已观察到的数据可以进行推理，从而作出最优决策[4]。贝叶斯公式如下：

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

其中，$P(h)$ 是在训练数据之前，假设 h 拥有的初始概率，称

基金项目：国家自然科学基金资助项目(60673041)
作者简介：黄 旭(1977－)，男，硕士研究生，主研方向：网络技术，信息安全；朱艳琴，教授；罗喜召，讲师、博士研究生
收稿日期：2007-12-22 **E-mail**：huangxu_sd@163.com

维普资讯 http://www.cqvip.com

为先验概率；$P(D)$ 表示将要观察的训练数据 D 的先验概率，即没有确定某一假设成立时 D 的概率；$P(D|h)$ 表示假设 h 成立的情况下，观察到数据 D 的概率；$P(h|D)$ 是给定训练数据 D 时，h 成立的概率，它反映了训练数据 D 的影响，称为后验概率。

2.3 影响分类实时性的主要问题

影响分类效率的因素主要包括中文分词和特征空间维数过高。由于中文单词之间没有显式的分隔标记，因此在处理时必须先进行分词。目前中文分词的基本算法主要有智能切分和机械切分 2 类[5]。智能切分以基于符号规则的人工智能为基础，复杂度高、实现难度大。机械切分可分为有词典和无词典 2 种，有词典切分算法的精度高于无词典切分算法。分词是一个复杂、耗时的过程，本文设计了无须对中文单词进行分词处理的分类系统，提高了分类速度。

高维特征向量处理的计算复杂度极高，不适于实时性要求高的应用，应采用降维手段以提高分类效率。文献[6]提出只采用频率较高的部分词构造文档的特征空间以降低特征空间维数。实验结果表明，当选择 1%的高频词时，分类效果最好。每篇文档只保留 50 个～100 个词可基本满足分类的需要，不会对分类结果产生影响。本文根据此观点优化了权重的计算方法以适应降维处理。

3 系统实现

本文设计的实时文本分类系统整体框架如图 1 所示。

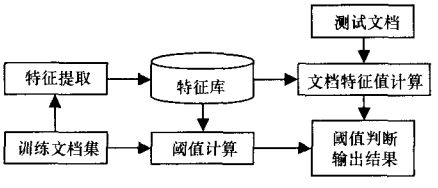

图 1 实时文本分类系统整体框架

在此方案中，训练集划分为正例集和反例集 2 个子集，可形式化为 $C=\{C_1,C_2\}$。其中，与国家安全敏感信息相关的文档构成正例集 C_1；其他文档构成反例集 C_2。实例集主要用于生成特征库并确定阈值。在训练过程中，对训练集中的所有文本进行特征提取，生成该类文本的特征库并计算阈值。在分类过程中，根据特征库计算测试文本的特征值，与事先确定的阈值进行比较并输出结果。由于训练过程可以进行非联机处理，因此整个系统的实时性主要体现在分类过程中，关键是提高文档特征值计算的实时性。此系统主要包括特征提取、文档特征值计算、阈值计算等模块。

3.1 特征提取模块

特征提取模块对训练集文档进行特征提取。先统计各单词出现的次数、单词总数、包含各单词的文档数等，然后计算各特征项的权重。特征项 T_i 在实例子集 C_j 中的权重 w_{ij} 可以用以下公式计算[7]：

$$w_{ij}=\frac{tf_{ij}\cdot df_{ij}}{\sqrt{\sum_{k=1}^{n}(tf_{kj})^2(df_{kj})^2}} \qquad j=1,2$$

其中，tf_{ij} 表示特征项 T_i 在该子集中出现的频数，是该特征项出现的次数与特征项总数的比值；df_{ij} 衡量特征项在该子集中的分布程度，用该子集中包含特征项 T_i 的文档比例来表示。按此方法分别计算 C_1、C_2 两个子集中所有特征项的权重。

为实现上述设计目标，本方案为特征项定义了 3 个属性，即 termName, filename, times，分别记录特征项名称、所属文件名以及在该文件中出现的次数，并在 SQL Server 2000 数据库中建立与之对应的 invertedTable 表。根据该表的数据，生成 wordsTable 表，其中包含 word, wordAmount, fileAmount, tf, df, w 等字段，分别记录特征项名称、出现次数、分布文件数以及据此计算得到的 tf，df，w 等值。2 个子集中的数据采用不同的数据表记录。

特征项 T_i 可能同时出现在 2 个子集中。为了进行类别判断，当特征项 T_i 出现时，应考查其所在文档属于类别 C_1 的概率，即 $P(C_1|w_i)$，用贝叶斯公式进行计算。根据 2.3 节的分析，为了提高分类速度，可以只取 C_1 子集中权重最大的若干个特征项作为目标文档类的特征表示，在本系统中可人为设定特征项数目，以所选特征项作为目标文档类的特征库。

3.2 文档特征值计算模块

对于一个实时分类系统而言，分类过程的速度决定了分类器性能，因此，应尽可能简化分类过程。因为文档特征计算是该过程的核心模块，所以不在该模块中对待分类网页做分词处理，而是在预处理后直接根据特征库进行匹配，从而有效提高分类速度。本方案统计匹配成功的特征项数目，设有 $w_1,w_2,\cdots,w_n$ 共 n 个特征项匹配成功，其中，下标 n 是标记特征项的序号，与上述特征库采用的下标含义不同，n 在数量上与特征库的特征项数目没有必然联系。本方案采用二维表记录待检测文档中的特征项及其出现次数，易于实现。根据联合概率公式计算当待分类网页中出现上述特征项时，该网页属于类别 C_1 的概率，即文档特征值。文档 i 的特征值用符号 P_i 表示。

3.3 阈值计算模块

本方案利用已生成的特征库计算正例集中的文档，求出每篇文档的特征值 P_i，并根据如下公式计算阈值 T：

$$T=\min\{P_i|i=1,2,\cdots,s_1\}+\delta$$

其中，s_1 为正例集 C_1 中的文档总数；δ 用于对阈值进行微调。不同应用所需阈值不同。例如，监控涉及国家安全的敏感信息时，用户不希望出现任何漏分现象，但偶尔把正常网页错分为非法网页是可以接受的，这种情况下应使阈值 T 尽量小；在过滤垃圾邮件的应用中，用户不能容忍任何把正常邮件错划为垃圾邮件的现象，而偶尔漏分是可以接受的，这时应使阈值尽量大。

4 实验及结果分析

本文方案用于实验的数据来源于 Internet。共选 1 200 篇网页文档作为训练集，其中，正例文档 300 篇。测试集采用不同于训练集的 320 篇网页文档，其中，正例文档 50 篇。实验步骤如下：(1)测试分类过程中去掉分词步骤后分类速度的提高情况；(2)测试分类能力。

统计同一篇文档的分类过程运行时间 t_c 及其分词时间 t_s，计算 $\eta=t_c/(t_c+t_s)$。笔者对大量文档进行计算，η 值均低于 0.1，证明本文分类系统的处理速度得到明显提高。分类能力主要采用精确率和召回率等指标进行评价[8]。测试结果如图 2 所示，随着测试文档数目的增加，本文所选测试集的精确率和召回率呈增长趋势。精确率曲线出现的波动说明目标文档分布不均匀。测试结束时，精确率和召回率分别为 85%, 94%。实验结果表明，本系统在大幅度提高分类效率的情况下，保持了较理想的分类效果。

(下转第 92 页)

维普资讯 http://www.cqvip.com

流程全局优化计划(第 8 行)，执行时间基本一致，这充分说明了设计的有效性。

4 相关工作

目前，Web 服务标准中并没有详细讨论 Web 服务组合的容错问题，设计并完全实现了一个良好的容错处理框架的例子也十分稀少，国防科技大学的 WSFMF 容错框架[4]也只考虑到相对简单或者单一服务的情况。文献[5]提出容错 SOAP 的概念，利用 Web 服务复制和日志的方法实现 Web 服务的故障恢复机制。工业界也在致力于制订容错的 XML 消息通信标准，如 WSRM。文献[6]扩展 UDDI 标准，提出主动 UDDI 的方法，建立一种基于代理的 UDDI 复制机制。虽然扩展 Web 服务标准可以解决容错问题，但因为标准体系本身不断发展变化，并存在兼容性问题，所以现有的研究成果在容错实现方面受到诸多限制，可实施性较弱。

5 结束语

本文叙述组合服务执行时的容错机制，介绍可能出现的故障类型，提出针对这些不同级别的故障可以采取的相应处理策略。然而本文研究的主要是集中式结构，如何将集中式结构中的服务组合容错结构改良并扩展到分布式结构中去，这是下一步工作的方向。

参考文献

[1] Business Process Execution Language for Web Services Java Run Time(BPWS4J)[EB/OL]. (2004-04-13). http://www.alphaworks.ibm.com/tech/bpws4j.

[2] Pleisch S, Schiper A. Approaches to Fault-tolerant and Transactional Mobile Agent Execution: An Algorithmic View[J]. ACM Computing Surveys, 2004, 36(3): 219-262.

[3] Gao Chunming, Cai Meiling, Chen Huowang. QoS-driven Global Optimization of Services Selection Supporting Services Flow Re-planning[C]//Proc. of the 1st International Workshop on Process Aware Information Systems. Huangshan, China: [s. n.], 2007: 16-18.

[4] 孙海燕，王晓东. Web Services 容错管理框架——WSFMF[J]. 计算机工程与科学, 2006, 28(4): 7-9.

[5] Liang D. Fault-tolerant Web Service[C]//Proceedings of the 10th Asia Pacific Software Engineering Conference. Chiangmai, Thailand: [s. n.], 2003: 310-319.

[6] Jeckle M. Active UDDI——An Extension to UDDI for Dynamic and Fault-tolerant Service Invocation Web, Web-services, and Database Systems[Z]. 2002: 91-99.

(上接第 86 页)

5 结束语

本文的排序算法可以有效提高网络原创文章的搜索排名，对网络竞争的公平性及鼓励网络创新起到较大的促进作用。但本算法并不能取代现有的排序算法，不能单独使用，在实际应用中只能对现有排序算法的排名结果进行一定的修正，使其更有利于首发站点。

参考文献

[1] CNNIC 第十九次中国互联网发展状况统计调查报告[Z]. [2007-01-23]. http://tech.tom.com/zhuanti/CNNIC_Report19.html/.

[2] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web[Z]. [2007-06-12]. http://www-db.stanford.edu/~backrub/pageranksub.ps.

[3] Haveliwala T. Effcient Computation of PageRank[Z]. [2007-05-10]. http://dbpubs.stanford.edu/pub/1999-31.

[4] Farahat A, LoFaro T, Miller J C, et al. Existence and Uniqueness of Ranking Vectors for Linear Analysis[Z]. (2001-09-10). http://www.damtp.cam.ac.uk/user/jcm52/hits.pdf.

[5] 李向伟，曹　博. 时间参数在 HITS 算法中的应用及改进[J]. 兰州工业高等专科学校学报, 2006, 13(2): 19-22.

[6] 常　璐，夏祖奇. 搜索引擎的几种常用排序算法[J]. 图书情报工作, 2003, (6): 74-77, 92.

[7] Nutch Version 0.8.x Tutorial[Z]. [2007-05-10]. http://lucene.apache.org/nutch/tutorial8.html.

(上接第 88 页)

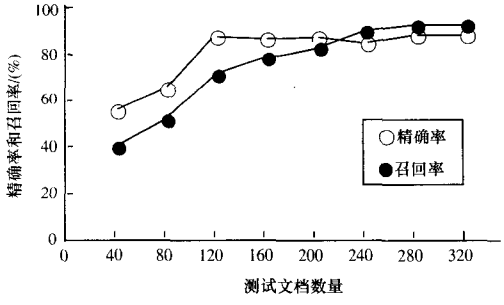

图 2　精确率和召回率测试结果

5 结束语

文本分类方法是文本处理中的热点问题，本文针对互联网信息监控处理量大、实时性要求高的特点，设计并实现了一种可用于实时场合的文本分类系统。加强类别特征提取、文档特征计算等是下一步研究的重点。

参考文献

[1] 郭　莉，张　吉，谭建龙. 基于后缀树模型的文本实时分类系统的研究和实现[J]. 中文信息学报, 2005, 19(5): 16-23.

[2] 冯是聪，张志刚，李晓明. 一种中文网页自动分类方法的实现及应用[J]，计算机工程, 2004, 30(5): 19-20.

[3] 王　斌，潘文峰. 基于内容的垃圾邮件过滤技术综述[J]，中文信息学报, 2005, 19(5): 1-10.

[4] Tom M M. 机器学习[M]. 曾华军，张银奎，译. 北京：机械工业出版社, 2003.

[5] 代建英. 汉语自动分词系统的研究与实现[D]. 重庆：重庆大学, 2005.

[6] 刘远超，王晓龙，徐志明，等. 文档聚类综述[J]. 中文信息学报, 2006, 20(3): 55-62.

[7] 李东艳. 互联网信息内容安全过滤方法研究[D]. 太原：山西大学, 2004.

[8] Kim H J, Shrestha J, Kim H N, et al. User Action Based Adaptive Learning with Weighted Bayesian Classification for Filtering Spam Mail[J]. Lecture Notes in Artificial Intelligence, 2006, 43(4): 790-798.