图书情报工作
第57卷 第14期 2013年7月

面向新闻领域的中文文本分类研究综述*

■ 薛春香 张玉芳

[摘 要] 在对文本分类及中文新闻分类概述的基础上,归纳出网络新闻文本特征及当前新闻文本分类特点,并总结新闻文本分类在新闻网站分类导航、话题识别与跟踪、个性化推荐三方面的应用。其后,总结中文新闻分类存在的问题,诸如缺乏通用语料和评价方法、分类体系粗略、分类维度单一等,并提出相应措施。最后,针对当前信息环境,提出新闻分类不仅将朝着多层次、多维度、跨语言方向发展,还将与多媒体信息、大数据、社会化媒体相结合。

[关键词] 新闻分类 文本分类 机器学习 中文信息处理
[分类号] G250 TP391
DOI:10.7536/j.issn.0252-3116.2013.14.022

1 引 言

新闻是“对新近发生的事实的报道”。新闻使用简练的文字概括了丰富的信息并频繁更新,且通过公开媒体传播。Web 2.0时代,网络丰富了新闻的来源,加速了新闻的传播。但面对爆炸式增长且杂乱无序的新闻,用户获取所需信息的难度增加,因此,迫切地需要对新闻进行有效的信息组织。

文本分类技术是信息组织、文本挖掘的重要基础,可以较大程度地解决信息紊乱的问题,帮助用户准确地定位所需的信息,被认为是处理海量信息的有力手段,近年来获得了快速的发展。随着中文信息处理的发展,中文文本分类得到了广泛关注。

将文本分类技术引入到新闻领域,代表性的应用有:Factiva依靠其智能标引体系实现数据库新闻信息的统一标引和分类[1];Google新闻使用计算机聚类算法识别出紧密相关的新闻报道[2];国内的新华社等媒体采用归类技术进行新闻分类[3]。归类和聚类的区别在于是否依赖已有的分类体系,相对于聚类结果的不确定性,归类根据待分文本特征与类别中对象特征的相似度进行分类,能够从整体的角度表现类别间的区别和联系,使得分类结果更具有系统性,因此,本文所研究的分类实指文本归类。

本文拟对相关研究进行概述,并分析新闻文本分类的特点和应用,指出现存的问题,预测未来新闻分类的发展趋势,以帮助新闻领域工作者了解新闻文本分类现状,促进新闻领域实现更加有效的信息组织和管理。

2 相关研究概述

2.1 中文文本分类概述

文本分类是按照预先定义的分类体系,根据文档的内容和属性,将文档集合中的每个文档归入到一个或者多个类别的过程。文本分类的发展历程如图1所示,当前已进入实用化阶段,被应用于众多领域,文献计量结果显示主要应用领域依次是信息检索、学习系统、数据挖掘、文本挖掘、模式识别等[4]。由于具有较可靠的理论基础和更好的分类结果,基于统计和机器学习的文本分类方法得到了学者的广泛关注,当前性能较好的机器学习方法包括K最近邻法[5]、支持向量机法[6]、类中心向量法[7]等。

相比国外对文本分类的研究,国内相关研究起步较晚。1981年,侯汉清教授对计算机在文献分类工作中的应用做了探讨并介绍了国外相关成果[8],带动了国内图书情报领域学者对于中文文本分类的研究。迄今为止,中文文本分类研究已经取得很大进展,也出现了较成功的系统,在北大天网、百度搜索等检索系统中投入应用,但同时存在一些难题,如中文预处理难度

* 本文系江苏省社会科学基金项目“数字报纸的自动标引研究”(项目编号:09TQC011)和教育部人文社会科学研究项目“电子报纸内容深加工研究”(项目编号:09YJC870014)研究成果之一。
[作者简介] 薛春香,南京理工大学信息管理系副教授,博士,E-mail:xuechunxiang@gmail.com;张玉芳,南京理工大学信息管理系硕士研究生。
收稿日期:2013-04-03 修回日期:2013-06-26 本文起止页码:134-139 本文责任编辑:易飞

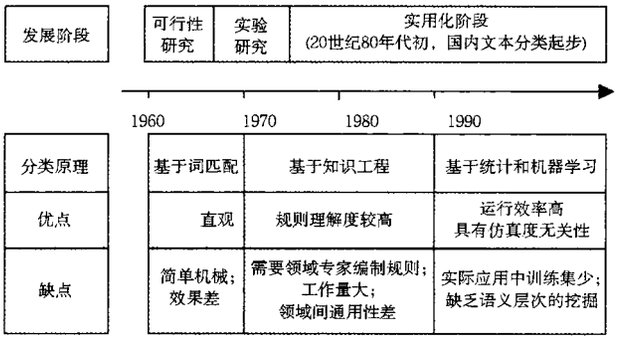

图1 文本自动分类的发展

大，缺乏公开的数据集、科学的评价体系等。

2.2 中文新闻文本分类概述

新闻信息是社会信息资源的重要组成，对新闻信息进行分类有助于实现新闻序化、新闻挖掘，进而指导决策等，意义重大。随着计算机技术和中文文本分类技术的发展，中文新闻文本分类在新闻资料分类的基础上逐步发展起来[9]，其发展历程如表1所示：

表1 我国新闻文本分类的发展

时间	新闻文本分类特点	重要成果
1979年及以前	各媒体运用互不流通的资料分类表	-
1980－1989年	新闻资料分类开始发展	新华社建立了数据库，编制了《新闻叙词表》
1990－1996年	各媒体的新闻管理全面计算机化	中国新闻资料学会研制了《中国新闻资料分类法》和《中国机检新闻资料分类法》
1997－2005年	各媒体交换电子数据，但分类标准不统一	国家科技部启动了国家科技攻关项目——《中国新闻信息技术标准》
2006年至今	网络新闻分类问题显现	国家标准化管理委员会发布了《中文新闻信息分类与代码》，杨丽英等人编制了突发事件新闻语料分类体系[10]

《中文新闻信息分类与代码》以政治、经济、文化作为分类大纲，采用主题为主、学科为辅的立类方法，促进了信息资源的无障碍交流和全社会共享。但是在科研领域，以此为分类标准的研究较少。当前中文新闻文本分类更多地是以新闻文本作为实验对象来判定分类算法的效果，在分类体系的选择上具有较大的随意性。在实际应用方面，广西日报社率先采用该标准开发了《广西日报电子版》数据库，实现了强大的新闻信息检索、加工功能，但是当前应用该标准的其他传媒机构还不多。

3 中文新闻文本分类研究

3.1 新闻文本的特征

从文本分类的角度分析，新闻具有以下两个特征：

3.1.1 新闻需要文本分类 首先，网络促进了新闻的生产，新闻量呈指数级快速增长，广大用户淹没在海量且杂乱的新闻中；其次，新闻的消费群体庞大，截至2012年6月，网络新闻用户规模达到3.92亿人，使用率为73%[11]；新闻具有广阔的应用市场；最后，新闻是通过公开媒体传播的，新闻中的信息更容易被忽略。

3.1.2 新闻分类具有可行性 新闻数量多、类别丰富，能提供好的训练集；新闻结构清晰，通常依赖固定模板，即使人工撰稿，也依赖一些固定结构，如新华体、华尔街日报体；新闻语言精练简洁，文章篇幅通常不大，新闻六要素能准确表述内容；根据新闻的结构和语言特点，优化特征提取，能获得更好的分类模型。

3.2 中文新闻文本分类的特点

基于上文提到的新闻的特征，将文本分类应用到新闻领域有重要实际意义。近年来已有不少这方面的研究，从这些研究中不难发现，区别于一般文本分类，新闻文本分类具有以下三个特点：

3.2.1 文本抽取结合网页特点 对新闻信息而言，文本抽取指的是将与新闻主题相关的文本如正文、标题等信息从网页中完整地提取出来。比起纯文本文件，网页表达的信息所包含的内容更丰富，学者也意识到文本抽取应结合网页特点。有学者提出对于新闻网页应该区分索引页和内容页，并在此基础上对网页进行去噪处理，进而进行文本分类[12]。蔡巍等[13]的研究提出了一种无需词典的运用网页页面特征从网页中抽取主题的实用算法，验证了文献[14]关于使用HTML标记有助于特征抽取的结论。如何从丰富的网页内容中有效地抽取出与新闻主题相关的文本是一个重要的研究区域。

3.2.2 文本表示考虑新闻特征 对新闻文本表示的研究中，学者在向量空间模型的基础上，根据新闻文本特点，进行了深入的研究。从特征项选择的角度，潘正高等人[15]参照付艳关于实体识别的加权方式，以新闻实体要素为特征构建了文本表示模型；魏程等人[16]的研究采用标题维度、正文维度、专有名词维度和时间维度的特征构成新闻的主体信息，构建了四维向量空间模型来实现文本表示；张永奎等[17]则提出在特征项中加入类别关键词。实验结果表明，采用上述文本表示方法后，分类性能都得到了明显提高。从特征项加权算法的角度，蔡华利等[18]考虑到HTML标签对词条的影响，改进了特征项加权算法；刘赫等[19]基于特征重要度改进了特征权重公式。充分分析新闻文本的特性，进而优化文本表示方法，有助于提高网络新闻的分类效果。

3.2.3　分类标准偏向主题而非学科　新闻分类区别于其他文本分类的另一个重要特点是类的划分偏向主题而不是学科。当前，国内外大型网络检索系统多采用指南型网络分类体系[20]，其大类设置采用以主题为中心或主题与学科结合的设类方式，相比传统的面向学科的分类体系，该体系不遵照特定的类目间排序规则，便于类目的调整和增补，因而更加灵活。由于面向主题的分类法更符合新闻的特征以及用户的查询习惯，在新闻分类的研究中也通常采用此分类法，例如马春华等人[21]的研究中采用了人民网的8个新闻类。也有学者应用专门的新闻分类体系，如张志平[22]所采用的是中文新闻信息分类体系，张永奎等人[17]采用了三层突发事件新闻分类体系，这些都是基于主题的分类体系。

3.3　中文新闻文本分类的应用

当前，中文新闻文本分类的应用主要集中于新闻网站分类导航、话题识别与跟踪、个性化新闻推荐三个方面。这三大应用的关系如图2所示，三者密切相关，且随着其关注点的升级，相应的应用等级不断提高。

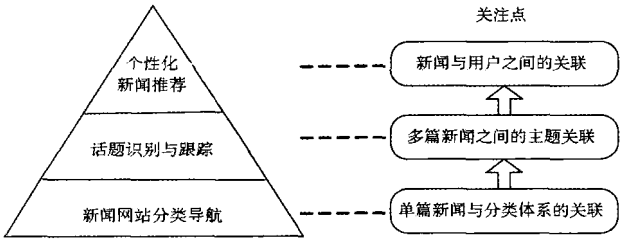

图2　中文新闻文本分类的应用

3.3.1　新闻网站分类导航　新闻网站以提供新闻、资讯等信息服务为主，其分类导航实现了将新闻资源按照一定体系组合形成信息集合，同时提供给用户分类体系的各级类目，以供浏览检索。艾瑞的《2012年第四季度数据产品监测报告》的数据显示：网站导航用户规模跃居第三，仅次于综合视频、网页搜索[23]。因此，研究中文新闻文本分类，提高分类的准确性，进而从内容上优化新闻网站分类导航的有效性，具有重要意义。

对于新闻网页的分类，不少学者倾向于使用性能较好的支持向量机算法，并在此基础上做了进一步改进：文献[24]着重比较了不同的权重计算方案与SVM的结合效果；文献[25]将遗传算法与SVM进行了融合；文献[26]对SVM算法中各种参数的设定进行了调优实验，这些改进都取得了更好的分类效果。

鉴于新闻网站已经成为突发事件最主要的传播媒体，针对突发事件的新闻网页分类也获得了较多关注：文献[27]在标引源和特征抽取过程中考虑了突发事件新闻文本结构的特点；文献[28]通过编程实现了按照事件类型和地理位置两种分类方式对新闻进行分类。

3.3.2　话题识别与跟踪　话题识别与跟踪(topic detection and tracking, TDT)是由美国国防高级研究计划署于1996年提出的，旨在开发一种新技术，以实现自动判断和识别新闻数据流的事件主题。TDT技术更强调发现新事件和关注特定事件的发展动态，被认为是一种特殊的文本分类过程，在信息安全、金融证券、行业调研领域有着广阔的应用前景。

TDT系列测评会议大大推动了相关研究的发展。近年来比较有创新性的研究包括：①对分类追踪器的改进研究，如改进向量距离模型追踪器[16]，使用在线主题模型[29]；②算法融合研究，如KNN和SVM算法多种结合策略的比较研究[30]；③语义信息的利用研究，如对特征词进行语义分组[31]，将新闻要素归纳为语义类[32]；④分类与聚类方法的结合运用研究[31,33-34]。

3.3.3　个性化新闻推荐　随着Web 2.0服务中RSS的广泛应用，新闻文本分类产生了新的应用，即个性化新闻推荐，用户能够从新闻分类结果中不同程度地自定义推荐内容，如Google网站的iGoogle、新浪的个性化推荐版块。通过分析用户所关注的新闻类别主动推荐相关的网络新闻分类结果，帮助用户及时获取感兴趣的新闻，将具有广阔的商业市场。近年来，学者对于新闻推荐系统的研究也在不断增加。根据推荐原理的不同，个性化新闻推荐可分为协同推荐、基于内容的推荐和混合推荐，其中，基于内容的新闻推荐系统的相关研究成果较多。近年来社交网络的出现推动了协同推荐的进一步发展，将两者结合起来进行混合推荐、实现优势互补，也成为近年的研究热点[35-37]。

4　问题分析

4.1　缺乏通用的标准语料和评价方法

国外文本分类技术研究起步较早，已建立了Reuters、20Newsgroups等标准的熟语料和统一的评价方法。而国内现有的分类标准虽已很详尽，但是缺乏对应的语料。现有的一些中文新闻语料库，如搜狗的中文新闻语料库、中国科学院自动化研究所的中文新闻分类语料库，所采用的分类体系都过于简单，不利于新闻自动分类的研究。而且在评价方法上，也还需要探索适用于中文的评价方法。

4.2　新闻噪声影响新闻分类质量

有统计称，网络新闻的真实性不足45%，常有新

闻标题与正文不一、正文前后表述不一、标题娱乐媚俗、夸张报道等现象。对失真的新闻做文本分类,文本表示环节会明显受到干扰,进而导致分类结果不具有实际意义,如一些新闻文不对题,在分类过程中会由于标题权重较高导致分类结果受影响;还于一些报道过于夸张,使得情感特征词的情感倾向被夸大,进而影响情感分类的结果。因此,在新闻信息抽取过程中,可运用网页分块技术识别标题块与正文块,并进行相关性计算,以确保新闻前后内容的一致性。运用LDA模型进行主题判定也是一个思路。

4.3 分类体系过于简单,不利于深度分析

当前网络新闻分类研究中的分类体系过于简单,多采取人为选定类别的方法,选定的分类体系类目少、层次少、类目间区分度大,趋于理想化。这与复杂的实际应用环境相背离。现实中,随着待分类新闻的数量增加,新闻的相似度增加,粗分类已不能满足用户需求,需要依赖更为科学的分类体系。依据多层次的新闻专业分类体系进行分类[38]以及提出的改进或重构传统文献分类法进而进行文本分类[39-40]的思路,值得借鉴。国家新闻信息分类标准《中文新闻信息分类与代码》采用层次编码法,对于多层次分类能够提供支持。

4.4 分类维度太过单一

当前网络新闻的分类多是从主题维度进行的。对于海量的新闻内容,应该提供多入口,实现分类的多维化。目前,已有从时间维度进行话题跟踪、从情感维度进行情感倾向性分析[41-42]、从地理位置维度进行分类[28]的研究,但综合多个维度的研究仍少之又少,这是未来的一个研究方向。

4.5 新闻专题平面化,缺乏深度

用户对于主题或事件的全方位认知需求推动了网络新闻专题的发展。新闻专题应该是一种深度报道,但是当前较多专题质量不高,通常只是相关信息的简单罗列堆积,虽然实现了信息的集成,却忽视了信息间的层次关系,缺少条理性和逻辑性,也缺乏系统性的梳理、归纳和总结,给用户冗余、杂乱的感觉。周科进[43]指出主题类和挖掘类是未来网络新闻专题的发展方向,认为这两类专题更具新闻价值。因而对网络新闻进行准确分类,进而进行深度挖掘以形成高质量、深度拓展的新闻专题,是网络新闻分类未来的一个重要方向。

4.6 跨语言新闻分类实现难度大

政府部门、商业机构需要把握全球最新的新闻信息,特别是信息安全、金融证券、行业调研等方面的新闻,从而支持政府部门或商业机构的决策与管理。但是这样的新闻集是多种语言的,语言成为信息共享的障碍,因此需要实现跨语言新闻分类。当前跨语言文本分类存在平行语料构建成本高、翻译软件系统使用受限且算法不可扩展、双语词典更新慢等问题,影响着新闻分类的效果,也制约着跨语言新闻分类的发展。基于词典的算法是实用性最高的,词典扩充的进步将推动跨语言新闻分类的发展。

5 总结和展望

中文文本分类是信息检索与数据挖掘领域的研究热点与核心技术,近年来得到了快速的发展。将中文文本分类技术应用到新闻领域,实现对海量新闻的有效组织,具有重要意义。本文在对中文新闻文本分类进行概述的基础上,对相关研究进行了梳理总结,并指出了现存问题,认为今后的研究方向主要有:

5.1 多层次新闻文本分类

当前文本分类多是扁平化分类,但是随着待分类文本的总量增加、类内文本相似性增加,在庞大的最底层类目中准确分类的难度大幅度增加,因此,充分利用分类体系的层次信息,采用逐层分类的思想进行多层次文本分类,能有效地降低分类算法的复杂度,同时保证分类精度,值得进一步研究。

5.2 多维度新闻文本分类

简单的新闻分类已不能满足用户的新闻需求,因此,需要从多个维度进行新闻分类,进而挖掘、整理、组织,以形成更有价值的新闻专题,从而满足具有不同关注点的用户的需求。

5.3 跨语言新闻文本分类

当前的分类研究通常是面向单一语种的,但是新闻信息全球化需要国际新闻媒体对不同语种的新闻进行组织和管理,如何通过发展跨语言文本分类技术解决此类问题,有待进一步研究。

5.4 多媒体新闻分类

如今新闻已不止于纯文字信息,而是更多地穿插了图片、音频、视频等其他媒体形式,因此,结合图片分类等其他媒体分类的研究,发展多媒体新闻分类,是未来新闻文本分类的一大趋势。

5.5 大规模新闻文本分类

随着大数据时代的到来,新闻文本也进入大规模状态,这是传统新闻文本分类没有遇见的问题。如何对大规模新闻文本实现快速高效的分类,是未来的一

个重要研究方向。

5.6　结合社会化媒体的新闻分类

Web 2.0 时代的用户不再是单纯地接受新闻，而是更多地参与到新闻的产生和传播中。利用社会化标签作为关键词辅助分类、利用大众标注进行分类测试等都是值得探索的方向。

参考文献：

[1] 李安. Factiva 新闻分类标引体系及其对我国的启示[J]. 图书馆建设，2003(3)：102－104.

[2] Google. Google 新闻的工作原理[EB/OL]. [2013－04－18]. http://support. google. com/news/ bin/ topic. py? hl = zh－Hans&topic = 2428790.

[3] 百度百科. 新华网[EB/OL]. [2013－04－18]. http://baike. baidu. com/view/154954. htm.

[4] 胡泽文，王效岳，白如江. 国内外文本分类研究计量分析与综述[J]. 图书情报工作，2011(6)：78－81.

[5] Yang Yiming. An evaluation of statistical approaches to text categorization[J]. Information Retrieval, 1999, 1(1－2)：69－90.

[6] Joachims T. Text categorization with support vector machines: Learning with many relevant features[M]. Berlin: Springer, 1998: 137－142.

[7] Lewis D D, Schapire R E, Callan J P, et al. Training algorithms for *linear text classifiers* [C]//*Proceedings of the* 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich: ACM, 1996: 298－306.

[8] 侯汉清，黄刚. 电子计算机与文献分类[J]. 计算机与图书馆，1982 (1)：5－14.

[9] 新华网. 我国新闻信息分类浅析[EB/OL]. [2013－04－13]. http://news. xinhuanet. com/new－ media/2006－02/10/content_4160298. htm.

[10] 杨丽英，李红娟，张永奎. 突发事件新闻语料分类体系研究[C]//中国中文信息学会第六次全国会员代表大会暨成立二十五周年学术会议论文集. 北京：清华大学出版社，2006.

[11] 第30次中国互联网络发展状况统计报告[R]. 中国互联网信息中心[2013－04－18]. http://www. cnnic. cn/gywm/xwzx/rdxw/2012nrd/201207/t20120723_32482. htm.

[12] 胡凌云，胡桂兰，徐勇，等. 基于 Web 的新闻文本分类技术的研究[J]. 安徽大学学报(自然科学版)，2010(6)：66－70.

[13] 蔡巍，王英林，尹中航. 基于网上新闻语料的 Web 页面自动分类研究[J]. 情报科学，2010(1)：124－127.

[14] Lim C S, Lee K J, Kim G C. Multiple sets of features for automatic genre classification of Web documents[J]. Information Processing & Management, 2005, 41(5)：1263－1276.

[15] 潘正高，侯传宇，谈成访. 基于命名实体的 Web 新闻文本分类方法[J]. 合肥工业大学学报(自然科学版)，2011(8)：1178－1182.

[16] 魏程，刘鲁，翟铭. 一种四维向量空间模型的 Web 新闻文本分类方法[J]. 微计算机应用，2010(3)：58－62.

[17] 张永奎，李红娟. 基于类别关键词的突发事件新闻文本分类方法[J]. 计算机应用，2008(S1)：139－140.

[18] 蔡华利，刘鲁，王理. 突发事件 Web 新闻多层次自动分类方法[J]. 北京工业大学学报，2011(6)：947－954.

[19] 刘赫，刘大有，裴志利，等. 一种基于特征重要度的文本分类特征加权方法[J]. 计算机研究与发展，2009(10)：1693－1703.

[20] 马张华，张宇萌. 指南型网络分类体系初探[J]. 大学图书馆学报，2000，18(3)：22－25.

[21] 马春华，朱颢东，钟勇. 结合新型文档频和二进制可辨矩阵的特征选择[J]. 计算机应用，2009(8)：2268－2271.

[22] 张志平. 基于“中文新闻信息分类与代码”文本分类[J]. 太原理工大学学报，2010(4)：402－405.

[23] 艾瑞咨询集团. 网站导航用户规模跃居第三[R/OL]. [2013－04－18]. http://search. iresearch. cn/scake/20130225/193568. shtml.

[24] Man Lan, Chew－LimTan, Hwee－Boon Low, et al. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines [C]//Proceedings of the 14th international World Wide Web Conference on Special Interest Tracks and Posters. Chiba: ACM, 2005: 1032－1033.

[25] 刘晓勇. 基于 GA 与 SVM 融合的网页分类算法[J]. 辽宁工程技术大学学报(自然科学版)，2010，29(5)：953－955.

[26] 张国梁. 专项主题新闻自动检索方法研究与应用[D]. 合肥：中国科学技术大学，2011.

[27] 王昌厚，罗永莲. 基于突发事件新闻网页的文本分类方法研究[J]. 长治学院学报，2006(2)：34－35.

[28] 郑魁，疏学明，袁宏永，等. 突发事件网络舆情信息分类方法研究[J]. 计算机应用与软件，2010(5)：3－5.

[29] AlSumait L, Barbara D, *Domeniconi C. On-line LDA: Adaptive* topic models for mining text streams with applications to topic detection and tracking [C]//*Proceedings of Eighth* IEEE International Conference on Data Mining. Pisa: IEEE, 2008: 3－12.

[30] 虞玲玲. 基于文本分类的话题跟踪及其一元语法模型的应用[D]. 南京：南京理工大学，2005.

[31] 宋丹，王卫东，陈英. 基于改进向量空间模型的话题识别与跟踪[J]. 计算机技术与发展，2006，16(9)：62－64.

[32] 刘炜，李明，杨合立. 基于本体的话题检测与跟踪技术[J]. 甘肃科技，2012，27(22)：42－45.

[33] 税仪冬，瞿有利，黄厚宽. 周期分类和 Single－Pass 聚类相结合的话题识别与跟踪方法[J]. 北京交通大学学报，2009(5)：85－89.

[34] 闵可锐，赵迎宾，刘昕，等. 互联网话题识别与跟踪系统设计及实现[J]. 计算机工程，2008，34(19)：212－214.

[35] 彭菲菲，钱旭. 基于用户关注度的个性化新闻推荐系统[J]. 计算机应用研究，2012(3)：1005－1007.

[36] 唐朝. 资源自适应的实时新闻推荐系统[J]. 计算机工程与设

计，2010(20)：4488－4491.
[37] Liu Jiahui, Dolan P, Pedersen E R. Personalized news recommendation based on click behavior[C]//Proceedings of the 15th International Conference on Intelligent User Interfaces. Hong Kong: ACM, 2010:31－40.
[38] Ha-Thuc V, Renders J M. Large - scale hierarchical text classification without labelled data[C]//Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. Hong Kong: ACM. 2011:685－694.
[39] Wang Jun. An extensive study on automated Dewey Decimal Classification[J]. Journal of the American Society for Information Science and Technology, 2009,60(11): 2269－2286.
[40] Waltinger U, Mehler A, Lösch M, et al. Hierarchical classification of OAI metadata using the DDC taxonomy [M]. Advanced Language Technologies for Digital Libraries. Berlin: Springer, 2011: 29－40.
[41] 徐军，丁宇新，王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报，2007,21(6):95－100.
[42] 陶富民，高军，王腾蛟，等. 面向话题的新闻评论的情感特征选取[J]. 中文信息学报，2010(3):37－43.
[43] 周科进. 网络媒体表现形式的集大成者:网络专题[J]. 新闻战线，2004(6):64－67.

Research Review on Chinese Text Classification in the News Field

Xue Chunxiang Zhang Yufang
Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094

[**Abstract**] Based on the review of text classification and news categorization, the features of news text and the characteristics of news categorization are concluded. The applications of Chinese news categorization on news site navigation, topic detection and tracking, and personalized news recommendation are summarized. Finally, this paper puts forward corresponding measures to solve existing problems about Chinese news categorization, such as low authenticity of news, idealization of classification system, and single dimension of classification.

[**Keywords**] news categorization text classification machine learning Chinese information processing

(上接第124页)

[7] 袁军，朱东华，李毅，等. 文本挖掘技术研究进展[J]. 计算机应用研究，2006(2):1－4.
[8] 吴根斌，丁振凡. 基于语义Web的搜索引擎研究[J]. 计算机与现代化，2012(8):130－133.
[9] 钱智勇，周建忠，贾捷. 楚辞知识库构建与网站实现研究[J]. 图书馆理论与实践，2010(10):70－73.
[10] 南通大学楚辞研究中心，南通大学图书馆. 楚辞研究数据库[DB/OL]. [2013－04－16]. http://zjyj.ntu.edu.cn/chuci/.
[11] 朝乐门，张勇，邢春晓. 面向开放关联数据的知识地图研究[J]. 图书情报工作，2012，56(10):17－23.

Study on Personalized and Related Retrieval Model Construction Based on Semantics: Taking Songs of Chu Research Database for Example

Tu Zhongqun Duanmu Yi
Library of Nantong University, Nantong 226019

[**Abstract**] This paper analyzes characteristics of semantic retrieval and mining of text concepts, and makes the semantic practice of Songs of Chu Research Database. Then it proposes a three－layer structure model of personalized and related semantic associative retrieval, which consists of ontology development, resource management and retrieval services and takes ontology repositories construction as core. This model integrates semantic dictionary, knowledge map, cross－database queries and project search as a whole, in order to get rid of the current restrictions of experiments and to promote the organizational development, retrieval and use of literature knowledge in related field.

[**Keywords**] semantic retrieval model construction personalized association text mining ontology the Songs of Chu