网络群体心理趋势智能分析模型研究

靳宇倡[1] 秦启文[1] 安俊秀[2]
(西南大学文化与社会发展学院 重庆 400715)[1] (成都信息工程学院软件工程学院 成都 610225)[2]

摘 要 网络群体是随着互联网网民的飞速膨胀而出现的虚拟聚集但真实存在的群体现象。由于互联网的隐秘性、自由性,使得网络群体能针对某一事物现象更真实地表述自己的观点。网络群体心理趋势分析是综合心理学、云计算、信息检索、自然语言处理、统计学等多学科理论和方法设计的智能模型。构建了以程序流为核心的网络群体心理趋势智能分析模型架构,突破了数据流方案,解决了云计算软件技术模式瓶颈。进一步研究了智能分析模型架构中核心模块的设计与实现,并以大学生网络群体的3万篇文本来检验该模型。结果表明,该模型能高效地实现网络群体心理趋势特征分析,能通过云图及结构方程模型给用户一个真实的感受。

关键词 网络群体,管道并行集成切词算法,哈希(Hash)散列算法,程序流,中文云图
中图法分类号 TP18 **文献标识码** A

Psychological Analysis of Trends Intelligent Network Model

JIN Yu-chang[1] QIN Qi-wen[1] AN Jun-xiu[2]
(School of Culture and Social Development, Southwest University, Chongqing 400715, China)[1]
(School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)[2]

Abstract Network groups are a virtual gathering but real group phenomenon with the rapid expansion of Internet users. Because of the Internet privacy and freedom, the Internet community can more really represent their own point of view aiming at a thing phenomenon. Psychological Trends Network is a comprehensive psychology, cloud computing, information retrieval, natural language processing, statistical theory and methods such as multi-disciplinary design of the intelligent model. Constructed the program flow at the core of the network group model of intelligent analysis of psychological trend in architecture, breaking the data stream programs to address the cloud computing model software bottleneck. Further researched the design and realization of the core module in intelligent analysis model structure. And the college students Internet Group's 30000 texts were used to test the model, the test showed that the model can efficiently realize the psychological trend of population characteristics of the network, give the user a real feelings through the cloud and the structural equation model.

Keywords Network groups, Integration of the segmentation algorithm for pipeline parallel, Hash algorithm, Program flow, Chinese word clouds

1 引言

互联网作为一种全新的高科技手段正以当初人们始料不及的惊人速度向前发展,它已成为信息技术浪潮的中心。据中国互联网络信息中心(CNNIC)发布,截至2009年6月底,我国网民规模已经达到3.38亿人,上网普及率达到25.5%。网络越来越深入人们的生活,网络的群体性逐渐凸显。相同兴趣爱好的人趋向群体化,形成网络群体,如专业论坛、社区、专业博客、QQ群、MSN讨论组等。他们通过论坛发帖、即时聊天、电子邮件等形式发表自己对某个问题的真实看法和观点,从而形成网络信息中庞大的发帖量。如百度贴吧中"迈克尔杰克逊去世"从2009年6月25日至2010年4月12日贴子数达4427570篇;"贾君鹏"从2009年7月16日至2010年4月12日贴子数达200814篇。从这些庞大的数据量中可以分析网络群体真实内在的群体心理特征、发展现状与成长趋势等具有潜在价值的信息。因此针对网络虚拟聚集、真实存在的特点,提出了网络群体心理趋势智能分析模型的研究与实现。

2 网络群体心理趋势智能分析模型架构

在深入研究云计算、搜索引擎、信息检索和网络群体特征的基础上,依据当前技术发展,提出了基于云计算的网络群体

到稿日期:2009-07-24 返修日期:2009-09-29 本文受四川省教育厅旅游专项课题(LY09-01),四川省青年科学基金(09ZQ026-068),国家自然科学基金(60702075)资助。
靳宇倡(1971—),男,博士生,副教授,主要研究方向为心理学研究方法、心里测量,E-mail:jinyuchang@gmail.com;秦启文(1955—),男,教授,博士生导师,主要研究方向为组织形象设计、组织文化和管理心理学等;安俊秀(1970—),女,硕士,副教授,主要研究方向为海量信息检索、移动计算及并行计算。

心理趋势智能分析模型架构，如图 1 所示。该模型架构以程序流为核心技术，由云信息层、网络群体心理趋势智能分析系统、用户查询框 3 部分组成；网络群体心理趋势智能分析系统由云采集层、云加工层、词频统计云图显示、群体心理趋势分析、云接口层、数据存储层云、云监控系统、云管理系统和调度系统组成。从架构上来说，它有如下技术突破点。

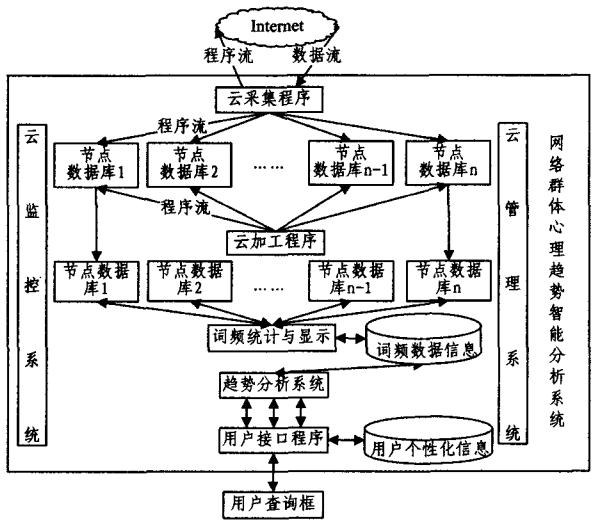

图 1　网络群体心理趋势智能分析模型架构

1. 以程序流为核心的网络群体心理趋势智能分析模型架构

云计算是并行计算(Parallel Computing)、分布式计算(Distributed Computing)和网格计算(Grid Computing)的发展。云计算强调的是后台庞大的数据处理问题，为了提高程序执行效率与系统的性能，仅仅采取并行执行方式已经不能满足处理海量数据的要求。因此在网络群体心理趋势智能分析模型中提出了采用程序流为核心的程序执行方式，如图 1 中的云加工程序，并编写了此模块程序代码，提高了执行效率。

采用数据流(Data Flow Diagram，简称 DFD)运行模式让庞大的数据在网络间流动已造成诸多瓶颈及系统运行效率低下等问题，如程序执行耗时长，数据易流失等。也就是说采数据流模式只适用于小规模数据，为了突破这种现状，在网络群体心理趋势智能分析模型中提出了采用以程序流为核心的工作模式，即将庞大的数据存储在节点数据库，让较小的程序流在网络间的流动来处理各个节点的庞大数据块，从而解决系统长期运行过程中的数据传输瓶颈，最终实现数据本地化，大大提高程序执行效率。不管是数据流还是程序流，都运用了流的最基本概念，即站点能够为用户提供速度足够快的数据和程序，以致用户或网络节点感觉到程序或数据似乎在当地机器上运行一样。

从图 1 中可以看出，Internet 上的海量数据经过爬虫，转化成数据流存储到节点数据库中，海量数据信息在节点数据库经过程序流的多重并行处理，最终得到网络群体心理趋势分析。

2. 分布式云数据库系统

以程序流为核心的网络群体心理趋势智能分析模型中运用了分布式云数据存储的机制，即把所有节点分成若干个组，每个组内包含若干个节点，组与组之间保持独立性。同时，在每个分组内，选择一些性能好、存储容量大、带宽高的工作站管理该分组内的其余工作节点，为检索分布化、并行化提供最可靠的支持。分布式云数据库系统采用了当今先进的开源的 Hypertable 分布式数据库。在 Hypertable 中，数据在存储前经过了排序和压缩，表中的数据类型都被串行化为字符数据。

3. 并行处理机制

以程序流为核心的网络群体心理趋势智能分析模型中，采用了并行计算的思想。该系统中采用分治法，主要运用在爬虫、切词、词频统计和中文文字云图中。

由于 URL 索引库中数据的存储采用的是分布式存储，因此在进行爬虫时可以并行执行，以大大降低爬取海量数据所耗费的时间。

在网络群体心理趋势智能分析模型中，采用了词典分块存储机制，从而实现并行地向多个字典块中进行匹配，若有一个匹配成功，则全部退出匹配过程。对一篇文章进行切词，也采用了并行执行，即可以从文章的不同部分同时切分，为了负载平衡，可以从文章的开头、1/3 处或 2/3 处并行处理。各个节点都完成了所分配文章的切词任务时，再将各部分归并拼接在一起，就可以用更少的时间完成整篇文章的切词。被切词的文章越长，效果越显著。

为了提高检索效率，各层之间实现分布式执行，每个层内部的功能模块实现并行执行，这个工作调度由云管理和调度系统来完成，云管理和调度系统就像一个管家程序，协调处理网络群体心理趋势智能分析模型的各项工作。云监控软件监控整个系统的“和睦相处”，如监视是否有对单个 IP 地址过分密集的访问工作。

3　云采集层设计

网络群体心理趋势智能分析模型中的所有数据都是通过一定规则让网络蜘蛛程序从互联网上爬取的。因此，在网络蜘蛛程序启动之前，系统中维护一个超链队列，它包含一些起始 URL，网络蜘蛛程序从这些 URL 出发，下载相应的页面。本文设计的云采集层系统主要模块包括：模拟 HTTP 协议功能、网页文件的编码转换模块、URLRank 算法模块、URL 提取算法模块、蜘蛛爬行状态的保存、多线程蜘蛛控制模块、蜘蛛程序流程控制、原始数据文件存储格式的定义等，如图 2 所示。为了提高效率，使用多个网络蜘蛛系统并行工作，让每个蜘蛛程序负责相应信息的数据爬取。为了便于将来扩展服务，网络蜘蛛程序应能改变搜索范围，当蜘蛛程序在一次爬行没有结束退出时，蜘蛛程序保存当前的爬行状态，当下次蜘蛛启动时可以继续爬取网页，从而提高网页爬行的效率和准确率。

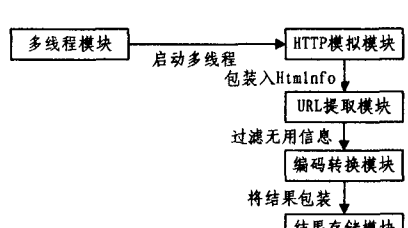

图 2　网络蜘蛛系统结构模块

在云采集层的设计中，主要对网络群体的特点进行了分析。网络群体是脱离了实体的限制，相对独立地存在于虚构的而又对应于现实的网络世界；网络群体有具体的社群形式，

如 BBS 社区、博客、OICQ 群等；网络群体之间的联系和互动是实质存在的；网络群体的参与者高度自主，个体间的交往突破了时空的限制；网络群体的构成具有扁平化、松散化、成员流动性大、道德约束力弱化等特有现象。

根据网络群体的特有现象，需要设计活动的超链队列，把某一群体的相关网址 URL 输入超链队列中。如我们要通过网络群体对大学生的择偶观念、学习态度和工作意愿等方面进行心理状况分析，那就要选取内容真实性程度相对比较高的大学生论坛、BBS 等网络资源，提取相应的 URL 输入超链队列。

针对大学生群体，要获取的信息来源是大学生的 BBS 和论坛，抓取 2008 年 6 月以来与大学生情感、学习、工作相关的帖子，存入相应数据库。这里我们选定了洛迦山水 BBS、同济网论坛、梦飞网 BBS、村里村外|中央民族大学论坛 BBS 站、中央财经大学论坛、中国政法大学 BBS、攀之苑 BBS 站等论坛，将其 URL 地址存入超链队列。接着利用蜘蛛程序爬取了 3 万篇以上文档。

4 云加工层设计

云加工层的主要功能包括：从原始网页数据文件中获取数据、分词算法定义、中文分词处理、英文分词处理、量词数字处理、标点处理、正排索引存储格式的研究等。云加工层的核心技术是中文切词算法。中文分词(Chinese Word Segmentation)技术属于自然语言处理技术范畴。本文首先对当前热门的中国科学院的汉语词法分析系统(Cascaded Hidden Markov Model，层叠隐马尔可夫模型)、Apache Lucene 自带的 3 种分词器 StandardAnalyzer，ChineseAnalyzer 和 CJKAnalyzer 以及对 Lucene 切词优化的版本：IK_Analyzer，PaodingAnalyzer(庖丁解牛)，Mmseg4j，MMAnalyzer 几种中文分词系统进行了深入研究。提出了针对网络群体心理趋势智能分析模型的管道并行集成切词算法，该算法是一种比较理想的切词方案。

网络群体所采用的网络语言与我们常规的文档有很大的不同。网络语言一般有移植借用、数字表意、缩写速配、符号传情四种社会文化生成方式，网络语言具有诙谐情感化、虚拟情境化和简约形象化等风格特征。网络群体中的日常交流用语有大虾或大侠——网络高手、斑竹——版主或网络管理员、网虫——沉溺网络的人、菜鸟——网上新手、白骨精——白领＋骨干＋精英公鸡、886——拜拜了、7456——气死我了、CU——See You、TMD——他妈的、3KU——thank you、^_^——表示男士温和礼貌的笑脸、：-D——表示开心地咧嘴大笑、洗耳朵——指听音乐及童语如“东东”指“东西”等等。这些如果用我们常规的切词方案，那是不可行的。因此在管道并行集成切词算法中主要对两部分进行了研究与实现——词典机制及分词机制。最后生成正排索引文档，同时支持词典和学习功能。

1. 词典机制设计

通过对网络群体内容的研究，发现需要把带有否定意义的短语作为一个词来处理，如“不高兴”、“很生气”、“非常不满”等，这些短语在常规词典中不是词，而我们若要利用词频对网络群体进行心理趋势分析，更需要把它作为一个词来处理，所以网络群体的词典不能采用常规使用的词典，只能由我们逐渐积累构成。设置基本词典和动态新词词典，基本词典包括国名词典、地名词典、机构名词典、产品名词典、商标名词典、简称词典、普通词典等。每个词典设有不同的权值，如国名词典比普通词典的权值高。词典中的词按照一定的算法，也划分相应的权值及词性，并且要根据每周搜索率的不同自动更新权值。动态新词词典主要是针对网络群体中频繁的产生新词而设置使用。这两种词典都采用了词淘汰机制，使得词典不会因不断添加新词而变得无限庞大。其中基本词典设置进入新词的阈值较高，这样可以保证基本词典的稳定性；新词词典设置进入新词的阈值较低，使得网络群体产生的新词能快速进入新词词典，以保证新词能被快速切分，从而更有效地实现网络群体心理趋势分析的准确性。

2. 分词机制设计

分词算法将国名识别、地名识别、人名识别、机构名识别、产品名识别、商标名识别、简称识别、省略语识别、数词识别、时间识别、字符串识别(邮箱)等均实现模块化，提供相应算法。

管道并行集成分词算法是对分词模块实行并行执行，不分先后，把所有可能出现的词切出来，其中所有的词都需要设定相应的权值及词性，再根据多种结果找出最佳词组合。

对歧义的处理，采用多策略交叉消歧匹配算法，它是一种能够检测句子中所有交叉歧义的中文分词算法，用 C 语言写成，专注于中文切词。该算法基于“长词优先”的切分原则，解决了切分路径数随句子长度的增长而呈几何级数增长的问题。算法的运算复杂度为 $Q(N)$，N 为句子长度。

对于新词，是重点研究对象，因网络群体的文档更易产生新词。采用基于规则的新词识别是不科学的，因此设计了以两次统计为主的新词识别算法。通过大量研究与实验，发现采用两次词频统计的方法得到新词的准确率基本为 100%。为了得到这样的准确率及高效性，采取两次词频统计得到准确新词，并进入相应基本词典中。采用此方案，需要建立动态新词词典。所谓动态新词词典是动态的增加与自减负的一个词典。建立新词词典的过程为：将串频统计的多字串的频度与较低的阈值(通常这一阈值由测试得来)做比较，判断是否在该阈值中，如果满足这一阈值则进入新词词典，并且在用户的访问量非常低的情况下，将无用的临时新词自动释放，同时也将经常访问并达到某一较高阈值的临时新词作为词语，加入基本词典。

没有最好的中文分词算法，只有更适合某一领域的中文分词算法，管道并行集成切词算法更适于在网络群体文档的切词中使用。

5 词频统计与云图显示

5.1 词频统计

对正排索引文件进行词频统计，统计出来的相关数据会生成词频统计表和关键词表存入数据库。词汇统计表如表 1 所列，关键词表如表 2 所列。

表 1 词频统计表(T_WORD)

逻辑字段	列名	数据类型	可否为空	备注
词语编号	ID	INT	NOT NULL	词语的编号
词语	WORD	varchar(500)	NOT NULL	词语
词频	TIMES	INT	NULL	该词在所有文档中出现的频率

所属类别	CLASS	varchar	NULL	该词属于大学生分析哪一方面
状态	B_STATE	INT	NULL	该词是不是已被选为关键词

表 2　关键词表(KEYWORDS)

逻辑字段	列名	数据类型	可否为空	备注
关键词编号	ID	INT	NOT NULL	关键词的编号
关键词	WORD	varchar(500)	NOT NULL	关键词
出现频率	TIMES	INT	NULL	该词在所有文档中出现的频率

对网络群体的文档进行词频统计后，词频最高的往往对网络群体特征的分析是没有用的，因此为了保证网络群体特征分析的准确性，采用了将前 500 个到 800 个词以表的形式显示给相关的心理学家进行选择(用户也可看到，也可进行选择)。接着根据心理学家所选择的词汇，再次对相关文档进行分词操作，统计出每篇文档相关词所出现的频率，以获得再测信度，如表 3 所列。这样我们才可使用 SPSS 对其进行心理评估。

表 3　每篇网页词频统计表(UN_URL)

逻辑字段	列名	数据类型	可否为空	备注
文档编号	ID	INT	NOT NULL	用来标识一篇文档
文档链接地址	VC_URL	varchar(500)	NOT NULL	记录该文档的链接地址
文档的总词汇量	ALLWORD	INT	NULL	记录一篇文档的总词汇数量
关键词总数	KEYWORD	INT	NULL	记录该文档中所选关键词总共出现次数
关键词 1	WORD1	INT	NULL	词 1 在文档中出现次数
关键词 2	WORD2	INT	NULL	词 2 在文档中出现次数
…	…	…	…	…
关键词 n	WORDn	INT	NULL	词 n 在文档中出现次数

5.2　词云显示

在研究搜索引擎技术的基础上，运用文本预处理、中文分词、统计词频等技术研究实现了中文文字云显示。即根据词频的大小来显示文字，词频越大，在画布上显示的就越大。同时还使用了字体多样性显示风格(由 Java API 提供)与显示方式多样性的技术，使得中文文字云可以产生形状与风格迥异的图片，在显示上实现了动态生成效果，拥有相对成熟的智能显示能力，从而实现快速理清网络群体主题和关键信息。本方案采用哈希(Hash)散列算法实现中文文字云显示。在椭圆方程$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$中，当变量 a,b 在一定的关系下，符合方程的所有点都会落在椭圆内部。其中变量 a,b 代表椭圆的长半轴和短半轴，是随机生成的确定的长度，变量 x,y 是由长半轴 a 生成位于椭圆范围内的坐标点(x,y)。即当变量 a,b 在一定的关系下，符合椭圆方程的所有点(x,y)都会落在椭圆内部。根据此方案，将所有要显示于画布上的词以这一坐标为中心随机生成在椭圆里，这样可以固定生成的词能够显示为一个云的形状，从而云的规则图形得以解决。为了使云图文字不出现重叠，采用 Hash 散列方法，散列方程利用由 random 生成的半轴长 a，随机取得的符合数学方程的点与词的大小存放在 Hash 表的一个节点上，下次生成的时候，通过计算周围词语中心点与词的长宽来判断是否可以放到这个点上，如果不行，重新生成，直到生成符合要求的点为止。

词云软件开发完成后，用户就可以运行该软件，这时会出现有 3 个输入框的页面，可以按照要求输入要查看的网络群体词频统计文档，单击“确定”，就可以得到相应的云图了，如图 3 所示。如果觉得它不够漂亮，只需单击重新建立，就会再生成一幅云图，直到用户满意为止。

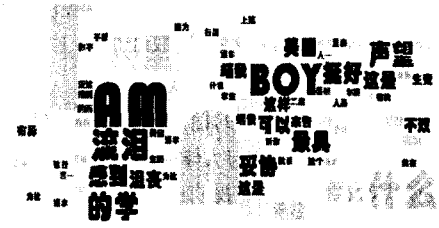

图 3　中文文字云的显示结果图

6　网络群体趋势分析

当某类网络群体的文档信息词频统计结束后，就可以通过心理学的相关知识，使用相关的统计软件，如 SPSS, SAS, 进行相应的探索性因素分析(exploratory factor analysis, EFA)，在此基础上，使用结构方程模型(Structural Equation Modeling, SEM)进行验证性因素分析(Confirmatory factor analysis, CFA)。如在对大学生网络群体择偶观念的研究中，通过管道并行集成切词算法，计算出大学生网络群体择偶词频并归类如下。

(1)凤凰男：出身农村，发奋读书，过上城市生活，娶了孔雀女(城市女孩代名词)，改变老家命运。

(2)经济适用男：相貌一般，性格温和，无不良嗜好，职业地位高，收入稳定，家庭责任感强。

(3)牛奋男：牛一样忠诚，具有奋斗能力，有可靠的人格魅力，具有上进心，对家庭感情的执着。

(4)传统男：家庭责任心重，忍辱负重，有大男子主义，自私，眼界狭隘，得过且过。

然后，通过结构方程模型进行验证性因素分析，验证性因素分析的概念模型图如图 4 所示。

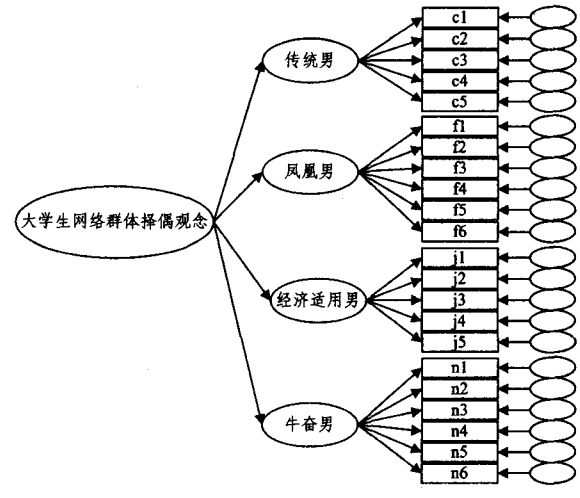

图 4　大学生网络群体择偶观念概念模型图

由此系统可以实现本土化的有针对性的分析，根据中国的具体情况、现代变化，分析出更加贴近、更准确的信息；及时

地发现中国大学生群体存在的问题，及时发出危机预警，以避免受错误方向的诱导和盲目跟从心理导致的弊端。

随着计算机技术的飞速发展以及网络群体带给人们观念的全新影响，对网络群体心理趋势的研究，将有助于更为准确地了解群体的心理状况。通过云计算技术、信息检索技术、中文文字云图技术、结合结构方程模型等现代统计技术，使得对于网络群体心理趋势的分析更为客观、有效。相信，通过信息检索技术注入心理词汇，可为心理学的研究注入新的活力，同时可使群体的心理研究更为科学、客观、准确。

参 考 文 献

[1] 张刚，刘悦，郭嘉丰，等. 一种层次化的检索结果聚类方法[J]. 计算机研究与发展，2008，45(3)：542-547

[2] 姜远，周志华. 基于词频分类器集成的文本分类方法[J]. 计算机研究与发展，2006，43(10)：1681-1687

[3] Westman H. *Finding your way in computer graphics*[J]. ACM SIGGRAPH Computer graphics，2006，40(2)

[4] Hsieh J W，Kuo T W，Chang L P. Efficient Identification of Hot Data for Flash Memory Storage Systems[J]. ACM Transactions on Storage，2006，2(1)：22-40

[5] *Chiang Y-J，Tamassia R. Dynamic algorithms in computational* geometry[R]. CS-91-24. Brown University，1992

[6] An Jun-xiu，Jin Yu-chang. Research on Text Classification Algorithm of Largest Dispersion Based on Term Frequency[C]// International Forum on Computer Science-Technology and Applications(IFCSTA 2009). 2009

[7] Haase C，Guy R. JAVA 动画、图形和极富客户端效果开发/Sun 公司核心技术丛书[M]. 北京：机械工业出版社，2008

[8] Law D. Taligent MVP in interactive statistical graphics [J]. Computational Statistics，2008，23：487-495

[9] Venugopal S. 数据结构从应用到实现(Java 版)[M]. 北京：机械工业出版社，2008

[10] 曹晓华，曹立人. 不规则几何图形识别取样特征的眼动研究[J]. 心理学报，2005，37(6)：748-758

[11] Kathryn E，Van Dam S A. Java 面向对象程序设计图形化方法[M]. 北京：机械工业出版社，2006，7：285-297

[12] 马亮，陈群秀，蔡莲红. 一种改进的自适应文本信息过滤模型[J]. 计算机研究与发展，2005，42(1)：79-84

[13] 王鹏，陈高云，安俊秀，等. 移动搜索引擎原理与实践[M]. 北京：机械工业出社，2009

[14] 王学松. lucene＋nutch 搜索引擎[M]. 北京：人民邮电出版社，2008

[15] 吴明隆. 结构方程模型－APOS 操作与应用[M]. 重庆：重庆大学出版社，2009：15-33

[16] 邱皓政，林碧芳. 结构方程模型的原理与应用[M]. 北京：中国轻工业出版社，2009：10-24

[17] 叶育鑫，欧阳丹彤. 语义 Web 搜索技术研究进展[J]. 计算机科学，2010，37(1)：1-5

[18] 申凡，钟云. 网络粉丝群体心理研究[J]. 南京邮电大学学报：社会科学版，2009，11(4)：27-31

[19] 吴太胜，陈业秀. 网络语言及网民群体的社会文化心理探析[J]. 广西社会科学，2007，9：171-174

[20] 尚得娟，张敏. 信息过滤系统中的混合式过滤算法[J]. 重庆工学院学报：自然科学版，2008，22(1)：118-121

(上接第 267 页)

从表 2 可以看出，当样本数较少时，算法 4 得到的结果与未进行规则约简得到的识别率大致相当，但耗费时间不到其一半；当样本数较多时(Abalone 与 Artificial Characters 数据集)，算法 4 的识别率超过其他两种方法，充分说明了算法的有效性。在时间消耗上，虽然稍大于基于模糊相似关系的规则约简算法，但识别率较后者提高较大。原因在于算法 4 较好地保持了原始决策表的特性，同时对冗余规矩进行了归并处理，在数据集较大时性能提升更为明显。

结束语 决策表是粗集进行知识获取的重要工具。而受客观世界及认知程度的影响，要获得完全精确、不包含任何冗余信息的决策表非常困难，因此在使用决策表进行知识获取前对规则进行约简是非常必要的。这不仅能去掉冗余信息，增强决策表对知识的描述能力，也能有效提高规则提取效率，促进粗集的进一步推广。

参 考 文 献

[1] Pawlak Z. Rough Set [J]. International Journal of Computer and Information Sciences，1982，11：341-356

[2] Pawlak Z，Grzymala-Busse J，Slowinski R，et al. Rough sets [J]. Communications of the ACM，1995，38(11)：89-95

[3] Pawlak Z. Vagueness-a rough set view [A]// Mycielski J，Rozenberg G，Salomaa A，eds. Structures in Logic and Computer Science：A selection of Essays in Honor of A [C]. Ehrenfeucht，Berlin：Springer-Verlag，1997：106-117

[4] Fernandez S J M，Murakami S. Rough set analysis of a general type of fuzzy data using transitive aggregations of fuzzy similarity relations [J]. Fuzzy Sets and Systems，2006，144：1-26

[5] Zhao S X，Wang X Z. Core and reduction from mutual relation view and their fuzzy generalization [A]// IEEE International Conference on Systems，Man & Cybernetics [C]. 2003：2611-2616

[6] 马玉良，杨家强，颜文俊. 基于模糊相似度的实值属性信息系统规则约简[J]. 浙江大学学报：工学版，2006，40(9)：1550-1553

[7] Li D Y. Artificial Intelligence with Uncertainty [M]. Beijing：National Defense Industry Press，2005：171-177

[8] Li D Y，Liu C Y. Study on the universality of the normal cloud model [J]. Engineering Science，2004，6(8)：28-34

[9] Li D Y，Liu C Y，DU Y，Han X. Artificial intelligence with uncertainty [J]. Journal of Software，2004，15(11)：1583-1594

[10] Li D Y. Uncertainty in knowledge representation [J]. Engineering Science，2000，2(10)：73-79

[11] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001

[12] Yuan S M，Cheng X Q. Clustering Method for Mining Quantitative Association Rules [J]. Chinese Journal of Computers，2002，23(8)：866-871

[13] Nguyen S H，Skowron A. Quantization of real value attributes-rough set and Boolean reasoning approach [A]// Proc. of the 2nd Joint Conf. on Information Science [C]. 1995：34-37