广播型网格的用户兴趣图谱

马建国[1],邢 玲[1,2],李幼平[3],文 丽[1]
(1. 西南科技大学信息与控制工程学院,四川绵阳 621010;2. 中国科技大学信息科学技术学院,安徽合肥 230026;
3. 中国工程物理研究院电子工程研究所,四川绵阳 621900))

摘 要: 用户兴趣图谱是实现智能代理与主动服务的基本依据. 而广播型网格的用户兴趣图谱有其自身的特点. 本文研究了 UCL (Uniform Content Locator) 在广播型网格中接收端的应用机制,建立了 UCL 解析、数据流控制和 UCL 应用机制,提出了建立基于 UCL 的用户兴趣图谱的方法,分析了信息的媒体结构与本体结构的特点,研究了相互间转换和在用户肖像模型建立过程中的作用. 利用实验建立的广播网格环境验证了该方法的有效性.

关键词: 网格; 广播; 用户肖像模型; 语义网; 智能代理
中图分类号: TN948. 61 **文献标识码**: A **文章编号**: 0372-2112 (2005) 01-0142-05

User Interest Spectrum of Broadcasting Gird

MA Jian-guo[1], XING Ling[1,2], LI You-ping[3], WEN Li[1]
(1. *School of Information & Control Engineering, Southwest University of Science & Technology, Mianyang, Sichuan* 621010, *China*;
2. *School of Information & Technology, University of Science and Technology of China, Hefei, Anhui* 230027, *China*;
3. *Institute of Electronics Engineering, China Academy of Engineering Physics, Mianyang, Sichuan* 621900, *China*)

Abstract: User Interest Spectrum (UIS) is the foundation of UCL (Uniform Content Locator) agent and active service in the broadcasting grid, which has many characteristics itself. UCL application mechanisms in the grid terminal are studied, such as the UCL resolution, the controlling of data stream, and the UCL application mechanism. At the same time, the method to build the UIS is put forward. Meanwhile, the medium framework and the ontological one of information are analyzed, including the characteristics, interchange and the effect of establishing the user profiles. Laboratory simulations demonstrate its feasibility and validity in the broadcasting grid.

Key words: grid; broadcast; user profile; semantic web; agent

1 引言

网格技术是未来信息共享的重要结构[1,2]. 与其它形式的信息网格相比,广播型网格有如下几个主要特征[3,4]:

(1) 传输网络无共享冲突

按照互联网的方式传播,一条信息要经过很多卡脖子的网关和路由器. 广播网格传输网络本质上无冲突、无瓶颈,因为它没有大量的网关和路由.

(2) 接收端广泛的用户群

接收端可以有无上限的用户群,不会造成"车多必然路堵"的现象,有效地缓解了带宽矛盾.

(3) 大幅度降低成本,消除数字鸿沟

广播成本除以海量的用户数目,每个用户只需承担微不足道的数额,信息服务的成本大幅下降.

(4) 消除语义鸿沟

信源端使用统一规范的 UCL 代码,用户在 UCL 代理的帮助下可以做到编者与读者的语义沟通.

(5) 消除信息垃圾

由于使用了 UCL 技术,从本质上杜绝了非法信息的传播,彻底消除了信息垃圾的滋生地.

因而,广播型网格真正能给用户提供时间、空间、内容的广泛 (pervasive) 整合的服务,容纳广泛的用户群体,真正做到运营的可持续发展.

文献[3,5]研究了广播型网格的基本原理与结构特点. 文献[6]研究了信源端 UCL 标引的方法和两级复用方法以及多映射与复用技术. 然而,广播型网格的应用研究是一个十分重要的问题. 要实现广播网格的智能代理与主动服务,用户兴趣图谱与用户肖像模型的建立是其关键技术. 本文依据广播型网格自身的特点,研究了在广播型网格中建立基于 UCL 的用户肖像模型的方法,研究了媒体结构与本体结构在模型建立过程中的作用,研究了建立用户图谱的机制,给出了基于 UCL 的用户肖像模型建立过程. 并在实验室建立的广播网格的环境下进行了实验研究.

2 UCL 在广播型网格接收端的处理

UCL 在以数据广播为核心技术的广播型网格发送端研究,如 UCL 协议框架建立、UCL 的映射、UCL 的标引、复用、传输等问题已经在文献[6]中详细讨论. 本文着重研究通过广播型网格传输后,在用户端 UCL 的解析、数据流控制、用户兴趣图谱生成和用户肖像模型的建立等问题.

2.1 UCL 在数据广播中的接收与控制流程

按照文献[6]中 UCL 的映射机制,α 映射、β 映射、γ 映射

收稿日期:2003-11-28;修回日期:2004-10-06
基金项目:国家 863 计划项目 (No. 2002AA121063);国家自然科学基金 (No. 60272014)

在接收端的作用如图 1 所示.

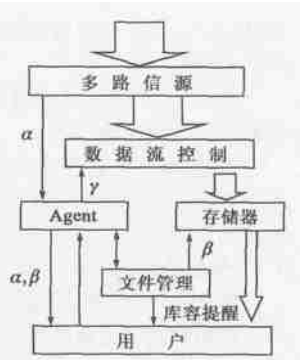

图 1 UCL 映射机制在接收端的工作流程

α映射是通过语义信道传输,形成电子节目单,是比较简略的 UCL 信息,以代码形式存在为主.β映射由文件头传输或者本地自动标引获得,是比较完整的 UCL 信息.γ映射是嵌在 TS 流中 21bit 信息,专供快速的数据流控制使用.因为广播网格传输的是信息量很大的数字媒体信息,快速反应的数据流控制可以大大降低对用户接收设备的压力.

接收端软件主要是对通过 CATV 网络发送过来的数据经数据接收卡后对收到的 IP 数据包和 UCL 信息进行识别、恢复、过滤、动态接收、选择存储、解析并进行镜像存储.

2.2　接收端信号处理的各个子模块

2.2.1　参数配置　参数配置主要包括接收卡初始化、复位、接收状态信息和调谐参数的配置.调谐参数主要包括频率、符号率、调制模式、和 PID 的设置.这些参数的配置主要是对硬件资源的管理.在接收数据之前,首先要对数据接收卡进行初始化,初始化成功才能接收到信号,然后进行调谐参数的设置,调谐参数设置的有效性关系到整个软件系统运行的可靠性.因此,调谐参数的设置在整个软件运行期间都起着至关重要的作用,在实验中我们采用的调制器的频率是 395000 KHz,符号率的范围为 0—7000 spbs,调制模式可选 16QAM, 32QAM, 64QAM, 128QAM 等不同的模式.这里 PID 的设置主要是将其添加到硬件接收卡上,以使接收卡能够识别此 PID,从而控制接收端下载数据.

2.2.2　UCL 解析　UCL 是整个系统的灵魂,它负责全面的信息管理工作.UCL 的解析自然就成为了本接收软件设计工作中至关重要的一部分.α映射传输 UCL 信息,类似于 DVB 标准的 SI 信息和 MPEG-2 标准的 PSI 信息,利用 UCL 信息可将其发送端的 UCL 信息映射成一个专用的电子节目播出表.用户通过该表可以浏览其发送端服务器上的文件的 UCL 信息,并可根据需要接收有用的信息,过滤不良的信息,也即体现了为用户主动提供信息的思想.

UCL 标引格式按照文献[6]的规定格式,其同步字节为 0×47,PID 共占 13 比特,连续计数器 3 比特文件号用来标识群中文件号总包和正在发送的包,用来进行解包处理,以使打包的数据从 TS 流中恢复出来.在发送端 UCL 信息是通过一个专用的信道来传输的,每一路的 UCL 信息通过 4 个 TS 包传输出去,并给一个固定的 PID.接收端通过该 PID 来识别和恢复 UCL 信息,并显示出来.在实验中采用一个列表框来显示解析出来的 UCL 信息.

2.2.3　数据流控制　由于在发送端打成 TS 流(188 字节)包的时候,已经插入了群、路和栏目等 UCL 信息,因此,在接收端可以根据 UCL 的三个字段来控制接收信息的群、路及栏目.实验中采用 WINDOWS 多线程机制控制数据下载,其中一个线程用来将数据取到缓冲区中,另一个线程用来读取和处理缓冲区中的数据.

2.2.4　文件管理　文件管理的主要功能是将接收数据的 UCL 信息与库存文件的信息进行全面的匹配,已库存的信息不再存储,否则需要存储或刷新.实验中,可通过对网站、题目、关键词、出版社、创作者等字段,迅速检索到本地磁盘上已经下载的用户感兴趣信息.

2.2.5　UCL 代理　UCL 代理(Agent)根据用户对 UCL 信息的选择从网络上下载(接收保存)用户需要的指定内容,同时统计用户对信息的阅读、需求频度,分析用户对信息的需求,主

表 1　基于媒体分类的语义地图

大类	栏　目												
新闻	国内	国际	社会	体育	科技	财经	娱乐	文教	聚焦	军事	图片	滚动	评论
财经	外汇	保险	期货	基金	股市	证券	银行	外汇	理财	专题			
体育	NBA	国际	国内	足彩	篮球	网球	综合	甲 A	英超	意甲	专题	女足	
科技	互联网	硬件资讯	软件天地	数码世界	通信时代	网络学校	论坛	IT	调查	评论	名牌		
娱乐	娱乐速递	明星	影视	流行音乐	综艺	流行时尚	下载	古典	图片	论坛	网上顶票		
游戏	论坛	网络游戏	攻略秘籍	电脑游戏	电视游戏	电子竞技	下载	联众	围棋	三国策	石器时代	天堂	空间互动
旅游	酒店预订	海外采风	国内线路	华夏纵览	旅游购物	旅游社区	旅游机构	自助旅游	论坛	游人游记	新马泰	出境旅游	机票预定
教育	文教咨询	外语天地	招生考试	留学移民	高考	考研	远程教育	校园	MBA	中考	书城	职业培训	校友录
健康	保健	健身	美容	心理	母婴	药品论坛	导医	家庭保健	健康商城	减肥	急救	两性	动态
短信	铃声	图片	订阅	彩信	言语传情	游戏	自写短信	点播	资费	帮助	下载	WAP	游戏
汽车	车迷世界	汽车行情	车文化	汽车名模	购车指南	新车报价	二手车	车主世界	品牌专区	车主之家	排行榜	维修保养	驾车出游
房产	楼盘展示	业主俱乐部	租赁市场	家居家装	装修论坛	楼市论坛	房屋租赁	二手房市场	装饰装潢	投资置业	地产精英	精选推荐	置业指南
女性	流行时尚	美容护体	情感世界	星座运程	品牌专区	精彩专题	角色女人	秘密花园	新闻人物	精致生活	情感婚恋	爱情快递	情书
招聘	求职宝典	职业动态	职业培训	英才推荐	最新招聘								
文化	文化专题	专栏	长篇连载	人气排行	观察	论坛	现场点击	摄影					
生活	休闲	玩乐	宠物	购物	心理	美食	笑话						

动为用户提供用户可能最需要的信息内容,完成对海量信息的粗选,并下载保存到用户本地硬盘中. Agent 可在本地硬盘存储的"海量信息"中实施精细的内容管理,把符合要求的课件送交浏览器显示,同时帮助用户实现智能化、个性化的信息浏览. 这样,智能代理就很好的起到一个信息代理、信息过滤的"绿色卫士"的功能.

UCL 代理包括下载控制、文件管理、基于本体结构与媒体结构的语义地图建立、用户兴趣图谱建立、用户肖像模型建立等模块.

3 信息资源的媒体结构与本体结构

UCL 是描述信息结构的重要依据. 在 UCL 的元数据结构中,有关于 UCL 的媒体结构表示与本体结构表示.

3.1 媒体结构与语义地图

定义 1 信息资源的整合与发布单位依据自身情况制定的信息资源的结构描述称信息资源的媒体结构或媒体分类.

如某网站定义的信息资源结构. 表 1 给出了新浪网的信息分类基本情况.

定义 2 信息资源依据某种分类结构而形成的可定位的数据结构称语义地图.

当然,不同信息资源发布单位发布的信息资源的分类是不一样的,而且也是不固定的. 媒体结构的混乱局面对于信息资源的交换和再利用都带来极大的不方便.

3.2 信息资源的本体结构

定义 3 权威部门或标准部门制定并规范的信息资源分类.

如国家技术监督局发布的《学科分类与代码》(国标 GB/T13745 - 92)收录 5 个大类 58 个一级学科 2500 个三级学科. 如:510.5025 通信-传输技术;510.5030 通信-网络技术. 其中,520 计算机科学技术学科的分类如表 2 所示.

对于图书资料的《中国图书馆分类法》(第四版)分为五大门类,二十二个大类,53,811 个类目(包括专用和通用类目).

表 2 基于本体分类的语义地图(520 计算机科学技术)

二级学科	三级学科及代码								
520.10 计算机科学技术基础学科	520.1010 自动机理论	520.1020 可计算性理论	520.1030 计算机可靠性理论	520.1040 算法理论	520.1050 数据结构	520.1060 数据安全与计算机安全			520.1099 计算机科学技术基础学科其他学科
520.20 人工智能	520.2010 人工智能理论	520.2020 自然语言处理	520.2030 机器翻译	520.2040 模式识别	520.2050 计算机感知	520.2060 计算机神经网络	520.2070 知识工程(包括专家系统)		520.2099 人工智能其他学科
520.30 计算机系统结构	520.3010 计算机系统设计	520.3020 并行处理	520.3030 分布式处理系统	520.3040 计算机网络	520.3050 计算机运行测试与性能评价				520.3099 计算机系统结构其他学科
520.40 计算机软件	520.4010 软件理论	520.4020 操作系统与操作环境	520.4030 程序设计及其语言	520.4040 编译系统	520.4050 数据库	520.4060 软件开发环境与开发	520.4070 软件工程		520.4099 计算机软件其他学科
520.50 计算机工程	520.5010 计算机元器件	520.5020 计算机处理器技术	520.5030 计算机存储技术	520.5040 计算机外围设备	520.5050 计算机制造与检测	520.5060 计算机高密度组装技术			520.5099 计算机工程其他学科
520.60 计算机应用	520.6010 中国语言文字信息处理	520.6020 计算机仿真	520.6030 计算机图形学	520.6040 计算机图像处理	520.6050 计算机辅助设计	520.6060 计算机过程控制	520.6070 计算机信息管理系统	520.6080 计算机决策支持	520.6099 计算机应用其他学科
.....									
520.99 计算机科学技术其他学科									

国际上常见的还有 DDC 杜威十进分类法、LCC 美国国会图书馆分类法和 UDC 国际十进分类法. 不难看出,目前的这些信息资源的本体结构大多是用于描述图书资料或者学科分类的代码. 对于现在占相当信息资源比例的网络信息资源并不太合适. 首先是网络资源包含非常广泛的形式,有文本、语音、视频、文件、程序等. 再者网络资源的变化十分迅速. 对照表 1 的媒体分类和表 2 的本体分类读者不难自己做出相应的判断. 因此,尽快制定适合网络信息资源的本体结构有着非常重要的意义.

3.3 信息媒体结构与本体结构的映射

一般用户所接触到的媒体结构远不止一种,因此要了解用户的兴趣不能单靠信息资源的媒体分类,建立基于本体分类的兴趣模型才更有意义. 首先要将媒体分类结构的信息映射到本体分类结构研究. 由于媒体分类结构的千差万别,加之又随时间而变化. 因此难以用普通的离散映射算法. 我们使用单隐层前馈神经网络分类器描述.

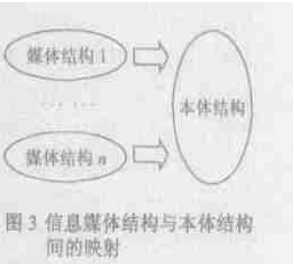

图 3 信息媒体结构与本体结构间的映射

在图 4 中 $x_i\ (i = 1, 2, 3, \cdots, n)$ 为 n 个媒体结构; $g_j\ (j = 1,$

2,3, …, *m*) 为 *m* 个隐层节点; d_k (k = 1,2,3, …, *p*) 为 *p* 个输出节点. 该映射器的构成有如下主要特点:

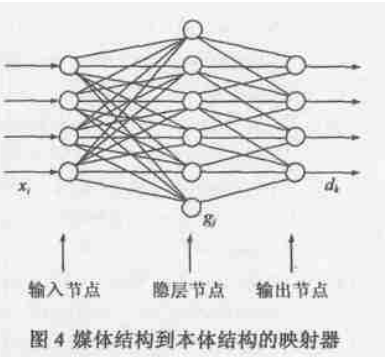

图 4 媒体结构到本体结构的映射器

(1) 本体分类结构一经确定,隐层节点数量决定着整个映射器的复杂程度. 可以利用不同媒体结构之间的共性,尽量简化关系,减少隐层节点数.

(2) 扩展媒体节点数不会大量增加运算工作量.

(3) 同一信息可能属于不同的本体分类,每个隐层节点可以有多个输出,但是不宜多于四个输出.

(4) 媒体结构的变化可以通过隐层节点的学习自动扩充隐层节点数来自动适应.

4　用户兴趣图谱

要实现对用户的主动服务,做到真正的个人智能代理,首先要对用户的阅读兴趣有相当的了解,并由此建立用户的肖像模型(profile).

4.1　用户兴趣图谱建立

要快速建立用户兴趣图谱,可对用户阅读兴趣的统计结果排序,截取高热度图谱形成如图 5、图 6 所示的"用户兴趣图谱". 根据前述媒体结构到本体结构的转换,可以将用户对信息的关注度在本体分类下进行统计,形成基于本体结构的"用户兴趣图谱". 基于媒体结构的用户兴趣图谱建立比较容易,中间不需要到本体结构的转换. 源于不同媒体的信息,其媒体结构不一致,将会造成兴趣图谱的不一致;因此,依据不同媒体结构建立的用户肖像模型不具有互换性.

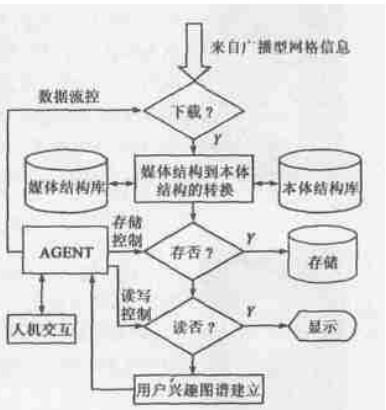

图 7 基于 UCL 的用户肖像模型

4.2　基于 UCL 的用户兴趣图谱的应用

基于本体结构的用户兴趣图谱机制在广播型网格的用户端的应用结构如图 7 所示. 从图 7 可以看出,兴趣图谱的形成是建立用户肖像模型的关键环节,是实现 UCL 代理与主动服务的关键环节.

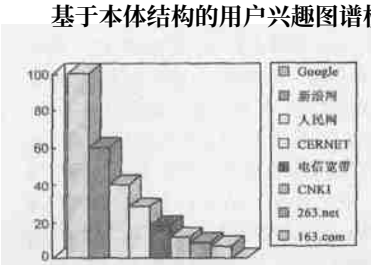

图 5 基于媒体网站的用户兴趣图谱

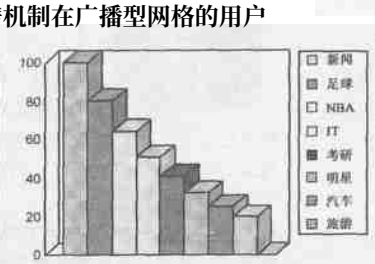

图 6 基于媒体栏目的用户兴趣图谱

5　实验仿真与结论

5.1　实验基本情况

依据上述用户肖像模型,我们在自己建立的广播型网格上进行了实验研究. 下面是实验广播型网格的基本情况.

设计规模:32 群 1024 路

传输信号总带规模:2 群 64 路　　调制方式:64QAM

传输协议在模拟频道的一个 8MHz 的带宽内,调制效率为 6bit/ band,最大可传输的比特率为:(余弦滚降系数为 1.15)

调制效率 ×模拟带宽 ÷余弦滚降系数

= 6 ×8MHz ÷1.15 = 41.7Mbps

传输内容 1:

http://www.sina.com.cn　　http://www.people.com.cn

http://www.edu.cn　　http://www.163.com

http://www.sohu.com　　http://www.263.net

等网站的内容经 UCL 标识后镜像传输.

传输内容 2:数据文件

传输内容 3:远程教育课件　50 路

5.2　实验内容

(1) 信源端的资源整合、标引、两级复用、映射、数据格式转换、64QAM 调制后到达 CATV-C 传输网.

(2) 在接收端使用了基于 PCI 接口的数据接收卡,在 WINDOWS2000 平台进行了 UCL 解析、数据流控制、语义地图的生成、基于媒体结构和本体结构的用户兴趣图谱建立、用户兴趣图谱的应用、人机交互等内容.


图 8 信宿端的 UCL 解析结果

5.3　实验结果

根据 5.1 和 5.2 的实验安排,我们以电子课件的接收、UCL 解析、用户兴趣图谱统计和用户肖像模型建立为例给出了实验的结果,如以下各图所示.

图 8 示出了在信宿端对某一群中的某一路的 UCL 的 α 映射的解析结果.

图 9 示出信宿端内容管理对话框的结构. 从图 9 可以看出,信宿端可以依据 UCL 方便地对文件及用户肖像模型进行管理.

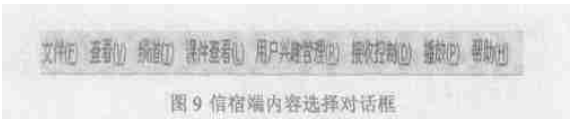

图 9 信宿端内容选择对话框

图 10 示出了用户根据 UCL 提供的信息对课件的搜索情况.

图 10 示出了用户根据 UCL 提供的信息对课件的搜索情况

图 11 给出了基于 UCL 的用户兴趣管理窗口. 用户兴趣图谱的统计图形显示方式见图 5、图 6 所示.

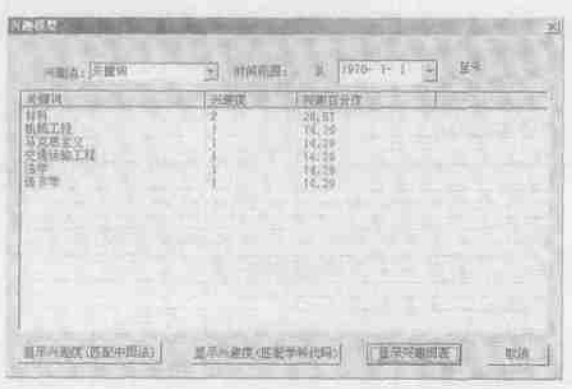

图 11 基于 UCL 的用户兴趣管理窗口

5.4 实验基本结论

(1) 利用本文提出的基于 UCL 机制对数据流快速下载的控制方法是可行的,TS 流中嵌入的 UCL 的 γ 映射对于快速下载控制是可行的,完全能满足目前普通用户终端对下载控制的要求.

(2) 利用 UCL 的 α 映射生成电子目录指导用户获取数据信息十分有效,而且 α 映射能够容纳大规模并播技术的多路 UCL 信息.

(3) β 映射的情况应分别对待. 对于结构统一规范的或者以文本信息为主的数据可以从本地自动获得 β 映射信息,符合国家《基础教育教学资源元数据规范 CELTS-42》的远程教育课件也容易从本地信息获取 β 映射信息. 数字音、影视频等难以自动标引的信息需要信源端的 β 映射信息. 当然,对于不需要详细了解数据 UCL 细节的可以省去 β 映射过程.

(4) UCL 信息对于用户端的人机接口、下载控制、兴趣图谱生成、用户肖像模型和主动服务的建立过程中始终起着关键作用.

(5) 用户兴趣图谱的研究是建立用户兴趣肖像模型和实现主动服务的关键技术.

5.5 实验中发现的问题

由于目前缺乏适合网络信息资源的本体结构描述规范,这给具有可交换的用户兴趣图谱和用户肖像模型建立带来极大的不便. 所以尽快制定这些规范是目前网络信息智能处理化处理的紧迫要求.

参考文献:

[1] Ian Foster. The grid: a new infrastructure for 21st century science[J]. Physics Today, 2001, 2(55): 42 - 47.

[2] Ian Foster. The grid: computing without bounds[J]. Scientific American, 2003, 4: 79 - 85.

[3] Ma Jianguo, Li Zaiming. The broadcasting grid[J]. Computer Science, 2004, 31(8): 5 - 7(Chinese).

[4] Li Youping. The second type network of information sharing[J]. Chinese Engineer Science, 2002, 4(8): 8 - 11(Chinese).

[5] Ma Jianguo. Information sharing technology with content indexing[D]. Doctoral dissertation. Chengdu: University of Electronics Science technology of China, 2004. 6.

[6] Ma Jianguo, Xing Ling, Li Youping, Li Zaiming. UCL indexing and transmission scheme in data broadcasting[J]. Acta Electronica Sinica, 2004, 32(10): 1621 - 1624(Chinese).

[7] Ma Jianguo, Li Youping, etc. Research of the national scale platform of distant education[J]. Distance Education in China, 2002, 7(186): 38 - 40(Chinese).

[8] Li Youping, Ma Jianguo, etc. The national scale platform of distant education[J]. Data Broadcast, 2002. 2(16): 1 - 5(Chinese).

[9] Tim Berners-Lee. The semantic web[J]. Science American, 2001, (5): 21 - 24.

[10] bert-Laszlo Barabasi, Eric Bonabeau. Scale-free networks[J]. Science American, 2003(5): 50 - 59.

作者简介:

马建国 男,1957 年出生于四川省梓潼县,获电子科技大学通信与信息系统专业博士学位,西南科技大学信息与控制工程学院教授,研究方向为信息系统技术,承担国家 863 计划项目与国家自然科学基金多项研究课题. 已出版著作三本,在电子学报等学术期刊发表学术论文五十余篇. Email: mjg_my@263.net

邢 玲 女,1978 年出生于四川省攀枝花市,中国科学技术大学信息学院硕士学位研究生,从事智能信息处理技术研究,已经在电子学报等学术期刊发表学术论文 8 篇.

李幼平 男,1935 年出生于福建省厦门市,中国工程院院士,中国工程物理研究院研究员,西南科技大学信息与控制工程学院院长,1957 年南京工学院无线电专业毕业,1957 至 1959 在清华大学无线电系研修多路通信与遥测,此后在成都电讯工程学院担任教师,1964 年 10 月,调往中国工程物理研究院,开始了核武器研究生涯,近年来在信息共享技术开展了研究,首先提出 UCL 和大规模并播技术概念,曾获得多种奖励,其中包括国家科技进步一等奖、国家发明二等奖、国防科技重大成果一、二、三等奖多项,1999 年获香港何梁何利基金技术科学奖,2000 年担任西南科技大学信息与控制工程学院院长,近年来开展了信息共享技术的研究,提出了大规模并播与 UCL 理念,在此基础上提出了国家文化网格与国家教育网格的思想.