

数据广播中的 UCL 标引与传输机制

马建国^{1,3}, 邢 玲^{1,2}, 李幼平^{1,4}, 李在铭³

(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. 电子科技大学通信与信息工程学院, 四川成都 610054; 4. 中国工程物理研究院电子工程研究所, 四川绵阳 621900)

摘 要: UCL (Uniform Content Locator) 是作者、编者和读者进行语义沟通的工具, 是数据广播中解决接收端数据信息的快速选择、信息过滤、智能代理和信息的主动服务的基础。本文系统地介绍了数据广播中 UCL 的概念, 提出了在数据广播中进行 UCL 标引的方法和两级复用方法, 研究了数据广播的 UCL-W 标引方案, 研究了多映射与复用技术, 最后给出了在数据广播的传输过程中, 发送端对数据进行采集、标引、映射、复用、格式转换和调制的工作流程。实验验证了理论方案的正确性和有效性。

关键词: 数据广播; UCL; 信息标引; DVB

中图分类号: TN943.6 **文献标识码:** A **文章编号:** 0372-2112 (2004) 10-1621-04

UCL Indexing and Transmission Scheme in Data Broadcasting

MA Jian-guo^{1,3}, XING Ling^{1,2}, LI You-ping^{1,4}, LI Zai-ming³

(1. School of Information & Control Engineering, Southwest University of Science & Technology, Mianyang 621010, China;
2. School of Information & Technology, University of Science and Technology of China, Hefei 230027, China;
3. College of communication & Information engineering, University of Electronics Science & Technology of China, Chengdu 610054, China;
4. Institute of Electronics Engineering, China Academy of Engineering Physics, Mianyang 621900, China)

Abstract: UCL (Uniform Content Locator) build a bridge between writer and reader for better understanding, which is a key technique for data processing such as receiving quickly, information filtering, service intelligently and actively in data broadcasting. The methods of UCL-W indexing and multiplexing are put forward for Web service in data broadcasting, and the techniques of multimapping and bistage multiplex are studies in this paper. And then the flow diagram of the data processing, including data gather, UCL indexing, multimapping, multiplexing, data format transform, modulation, are figured out. The experiment result shows that the creative methods and the new designs are correctly and efficiency based on DVB-C network.

Key words: data broadcasting; UCL; information indexing; DVB

1 引言

1948 年信息科学的奠基人 C. E. Shannon 在定义信息量时, 作了忽略语义的假设^[1], 因为当时世界上还没有出现“多信源对多信宿”的电子传媒, 也不存在“与谁通信”的困惑。今天不同了, 每人每天都要面对无力读完的海量的信息, 很需要获取有关“信息的信息”即语义信息。万维网 (WWW) 创始人 Tim Berners-Lee 认为^[2], 当今的互联网还只是一个脾气倔强的大孩子, 不可能完全实现他的初衷, 他领导的 W3C 开展了“语义互联网” (Semantic web) 的工作。

然而, 这种仅仅依托互联网一种网络的结构, 在网民、网页急剧增长之后, 出现了“带宽瓶颈”、“信息垃圾”、“数字鸿沟”等诸多问题, 要实现 Tim Berners-Lee 的理想并非易事^[3]。事实上, 随着数字电视逐步进入家庭, 承担传输任务的数据广播技术正越来越受到人们的关注。如何利用带宽资源十分丰

富的单向广播信道进行电子报纸、热门网站、股市行情、远程教育、数字图书馆等数据广播的增值业务; 如何解决接收端数据信息的快速选择与下载、信息过滤、智能代理和信息的主动服务等关键问题; 以至于如何解决万维网的“信息垃圾”、“带宽瓶颈”和“数字鸿沟”等棘手问题? 在数据广播中建立规范的信息标引与传输机制是一个十分重要和十分紧迫的问题。

文献[4]提出了“UCL” (Uniform Content Locator) 概念。近几年来, 一些研究机构、公司和高校开展了这方面的研究工作, 取得了一些局部的阶段研究成果。但是, UCL 的研究才处于研究的初期, 目前的研究缺乏统一的规划, 在信源端、网络传输、接收端没有统一的技术标准, 一些研究仅仅针对局部的应用, 对数据广播中的信息标引存在一些误区^[5]。本文系统地研究了在数据广播中信息标引的理论问题; 给出了数据广播中关于信息标引、复用、映射的基本定义; 建立了多映射模型; 试图从信源端的研究出发, 规范数据广播中的 UCL 元数据框架,

建立大规模并行广播与 UCL 的传输模型。

2 UCL-W 规范

2.1 UCL 定义

目前普遍采用的方式是将信息空间视为“按地址定位”的空间,确切地说,是按信息“存储源地址”定位的空间。如今在 Internet 上广泛应用 URL (Uniform Resource Locator:统一资源定位器)就是如此,而并不是真正的“资源定位”。

UCL 统一内容定位 (Uniform Content Locator) 是网络信息资源描述结构。

UCL 的目的是解决网络信息资源的发现、查找、识别、控制和管理问题。在信息空间 (cyberspace) 里,每一份多媒体文件都是一个多维矢量。矢量的模量(长度)是文件字段数,矢量的方向取决于对文件内容进行精细定位的一组代码,即 UCL 代码。UCL 代码对文件内容的类别、主题、出处、时段、作者、关键词、分类代码等作出多维度的标引。读者的需求和文件的内容都用 UCL 矢量来表达,通过对 UCL 矢量的关联计算,在浩瀚的信息空间中按内容准确定位文件。

设 UCL 的向量表示为

$$U = (u_1, u_2, \dots, u_n) \quad (1)$$

式中 $u_i, i = 1, \dots, n$ 是 UCL 的分量数,一般与被描述对象、应用领域、传输方式、用户终端形式有关。

为了区分 UCL 的描述对象与应用领域,可以定义应用于某些特定领域的 UCL。如用于描述 Web 信息资源的 U_w ,用于描述远程教育资源 (Education) 的 U_e ,用于描述数字视频 (Video) 的 U_v 。在后面的讨论中我们将会发现,URL 只是 UCL 的一个分量。

2.2 数据广播中 UCL 研究^[6,7]

在数据广播中研究 UCL 技术有其自身特点。首先,数据广播技术随着数字电视的推进而高速发展,因为有着巨大的经济利益,市场前景十分广阔。再者,在很多国家,广播是一种特殊的服务领域,事关国家安全问题。WTO 最后文件中电信

附则也作了相应的规定。因此,无论从国际法规的范围看,从数字电视与数据广播巨大的市场来看,还是从国家的信息安全等诸方面考虑,广播电视行业均有其自身的特殊性。所以,形成具有自主知识产权的标准是数据广播当前的十分重要和非常紧迫的问题。

UCL 的研究目标是要形成一个国家数据广播的元数据标准。也就是说,想要进入“国家数据广播网”的数字资源,都必须遵循 UCL 规范,在信息流程的适当环节进行 UCL 元数据标引,实现入网信息的有序化,最终将为数据广播终端用户的数据截取和本地交互使用创造极大的方便。因此,在国家标准这一层面,UCL 是统一、标准(强制性、法规性)的资源有序化准绳。

从传输层面上看,数据广播是以传输大流量的数字媒体业务,传输网络的带宽均超过数十兆。对于低成本数字接收终端的硬件要求来看,对数十兆 bps 数据进行快速的选择与下载,减轻大量数据对数据终端的压力也是一件十分有意义的工作。

UCL 是沟通编者与读者的工具。编者用它标引信息资源,用户用它表达阅读意向。用户 UCL 代码与文档 UCL 字段中元数据的匹配过程,实质就是在 UCL 标准下的信息检索过程;而匹配得到的文档方便地为用户所用,体现了信息的个性化按需服务。从这个角度讲,UCL 是信息时代的“绿色卫士”:把浪费人们精力的无关信息拒之门外,节省人的注意力资源。

2.3 UCL-W 基本规范

我们研究了基于 Web 的 UCL 结构。令其向量为 U_w , 则

$$U_w = (u_{w1}, u_{w2}, \dots, u_{wn}) \quad (2)$$

式中 $u_{wi}, i = 1, \dots, n$ 是 U_w 的分量数。

表 1 给出了我们在实验研究中使用的元数据规范框架。

UCL 的规范描述框架及 U_w 的字段定义考虑了如下原则:

- (1) 基本字段尽量与现行国际标准或工业标准一致。如 Language、Date、Format 等。

表 1 基于 Web 的 UCL 元数据规范框架

类属	元素名称	中文名	说明	举例	元素编码体系
资源内容	u_{w1} : Group	大类	信息资源的一级分类	体育	行业标准
	u_{w2} : Subject	栏目	信息资源的二级分类	足球	自定义
	u_{w3} : Title	标题	新闻标题	中国出线了!	文本描述
	u_{w4} : Key words	关键词	信息资源主题、内容的关键字或词组 (建议使用受控词汇)	世界杯、杨晨	中国主题词分类表、 汉语主题词表
	u_{w5} : Description	简介	资源内容的文本描述	文本描述
	u_{w6} : Language	语言	信息资源所使用的语言	简体中文	ISO 639-2
知识产权	u_{w7} : Creator	创作者	信息资源的制作人个人或组织	新华社	文本描述
	u_{w8} : Publisher	出版者	负责信息资源发布的实体	www.sina.com	URI
外部属性	u_{w9} : Date	日期	资源创建的日期或其它相关的日期	2003-06-25	W3C-DIF
	u_{w10} : Type	类型	资源的种类或形式 (如文本、图像、声音、软件、数据等)	text	DCMI Type Vocabulary
	u_{w11} : Format	格式	信息的数据形式、尺寸以及操作指示	html	IMT
	u_{w12} : Classification code	分类代码	资源的学科分类代码	如:510.5035	GB/T 13745-92 中图分类法
	u_{w13} : expansion	扩充	用户自定义	如:价格等	行业标准或自定义

(2) 不能使用国际标准或国家标准的可以借用行业标准或自定义. 如 Group、Subject.

(3) 为了适应数据广播中的快速处理,特别是用户端的快速下载的硬件处理,对 Publisher、Group 等进行高效率的编码. 要求这些字段应该十分精炼.

(4) U_w 的所有字段不是必需的,可以根据应用领域、传输方式、用户终端自由选择.

(5) 为了满足用户的特殊需求,可以扩充字段.

3 并播与复用技术^[8-10]

3.1 并播与复用的基本概念

为了有效利用数据广播的信道资源,在数据广播中引入竞争机制,充分调动 ICP 的活力,文献[3]提出了在数据广播中的“大规模并行广播”方法.

大规模并播技术 将数据广播的一个信道(在一个 8M 有线模拟带宽上传输数据广播信息)划分成若干的独立经营的子信道,从而形成的相互独立经营的竞争机制.

总带 数据广播的一个信道所包含的所有带宽资源.

子带 一个基本的带宽资源,如 32Kbps.

例如,以总带为 32Mbps,子带带宽为 32kbps 为例,

$$32\text{Mbps} = 4\text{MB/s} = 14.4\text{GB/时} = 345.6\text{GB/天} \quad (3)$$

如果把 32Mbps 分给 1024 个 ICP,每个 ICP

$$32\text{Kbps} = 4\text{KB/s} = 14.4\text{MB/时} = 345.6\text{MB/天} \quad (4)$$

每一个 ICP 都拥有 32Kbps 永不断线的常在带宽,经时间的累积,可以推送大量的数据进入用户终端.当然,由于 DVB 数据帧的控制字和 CRC 占用一些字节,传输过程还会有 2%~9% 的开销,实际传输的净荷略低于上述数字.

元带 用来传输各子带 UCL 元信息的专有信道,或称语义信道,带宽一般与子带相同.

群 包含一些子带的集合.我们定义为包含 32 个子带的集合,正好为 1Mbps.

路 ICP 独立使用的信道.可以为一个子带或 N 个子带.

复用 利用数据广播信道传输多个子带的方法.根据具体的实施方案,可以有级复用.

3.2 两级复用模型

根据现有数据广播的带宽,一般分为两级复用比较合理.两级复用模型如图 1 所示,均采用 TDMA 的多址方式.

如果每一路仅包含一个子带,则一个 32Mbps 的信道

可以划分成 32 群、1024 个子信道即 1024 路.

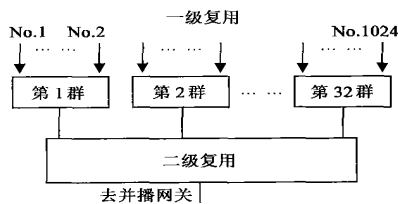


图 1 两级复用模型

4 数据广播中的 UCL 标引

4.1 UCL 标引基本概念

UCL 标引 从数据信息提取元数据的过程.

在信源端,应该对数据文件进行 UCL 的标引,以便将 UCL

信息连同数据信息本身复用、编码、调制、传输.

4.2 UCL 的标引方法

标引方法可以是自动标引与人工标引方法.

人工标引方法一般用填表方式.人工标引的特点是简单.但是,人工标引的方法不适用在线实时数据信息的标引.

自动标引方法是利用软件自动标引.其特点是快速,不需要人为干涉.其难点是对影视、音像作品难度很大.对数据信息,需要自然语义理解较为专门的知识.

目前最容易做到自动标引的是网页和文本.目前研究标引技术遇到的最大困难是网络信息的本体结构研究滞后.

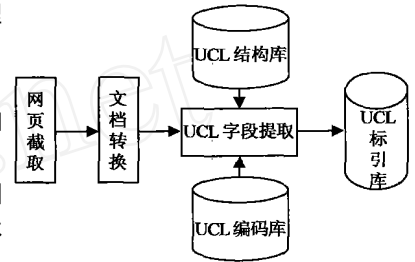


图 2 UCL 自动标引示意图

图 2 给出了网页的 UCL 自动标引

的示意图.图中的 UCL 结构库、UCL 编码库和 UCL 标引库均使用 XML 格式.

5 数据广播中 UCL 的传输与映射^[11]

5.1 UCL 映射的基本概念

UCL 映射 将已经完成标引的 UCL 信息进行某种变换,以方便某种传输和表示的需要.

5.2 UCL 多映射方法

为了使 UCL 信息在数据广播中的数据收集、内容分发、快速下载,以及 Agent 的全过程有效服务,应该对 UCL 采用多种映射方式,以分别满足不同用途和不同阶段的需求.我们在研究中使用了一、二、三种映射.

对于第 j 群 k 路的 UCL 向量可以表示为 $U_w^{j,k}$,则

$$U_w^{j,k} = (u_{w1}^{j,k}, u_{w2}^{j,k}, \dots, u_{wn}^{j,k}) \quad (5)$$

$j=1, 2, \dots, l; k=1, 2, \dots, m$. l 与 m 分别表示总带中群的数量和路的数量.

映射 向量 $U_w^{j,k}$ 在元带的像.定义为 $U_w^{i,k}$,则

$$U_w^{i,k} = (u_{w1}^{i,k}, u_{w2}^{i,k}, \dots, u_{wp}^{i,k}) \quad (6)$$

$u_{wi}^{j,k}$ ($i=1, 2, \dots, p$) 是按照映射法则的各分量, p 代表 $U_w^{j,k}$ 的分量个数.一般情况下,映射使用 1-1 映射法,则 $p=n$.

映射存在于元带中,在接收端表现形式为用户的实时的节目单.定义多群多路的在元带中的映射向量为 U_w .对于总带中含有 μ 群 μ 路的信道,有

$$\begin{aligned} U_w &= \{ U_w^{j,k}, j=1, 2, \dots, l; k=1, 2, \dots, m \} \\ &= (u_{w1}^{1,1}, \dots, u_{wp}^{1,1}, u_{w1}^{1,2}, \dots, u_{wp}^{1,2}, \dots, u_{w1}^{1,m}, \dots, u_{wp}^{1,m}, \\ &\quad \dots, u_{w1}^{l,m}, \dots, u_{wp}^{l,m}) \end{aligned} \quad (7)$$

映射 文件或者数据包的自身 UCL 标引方式.主要用于对文件的 UCL 信息管理和代理.其表现形式为 UCL 索引,存在于基于 XML 的 UCL 索引库.具有 n 个分量的在映射下的向量表示为

$$U_w^{j,k} = (u_{w1}^{j,k}, u_{w2}^{j,k}, \dots, u_{wq}^{j,k}) \quad (8)$$

在实验中,我们选取 $u_{wi}^{j,k} = u_{wi}^{j,k}, i = 1, 2, \dots, q, q = n$.

映射向量 $U_w^{j,k}$ 在数据链路层的像,用于接收终端的快速下载控制,具有 n 个分量的映射下向量 $U_w^{j,k}$ 表示为

$$U_w^{j,k} = (u_{w1}^{j,k}, u_{w2}^{j,k}, \dots, u_{wr}^{j,k}) \quad (9)$$

由于映射存在于数据链路层,为了保证数据传输效率,映射要求高效精简.我们在数据广播网的传输实验中,取 $n = 3$,且 $u_{w1}^{j,k}$ 代表的是群编号, $u_{w2}^{j,k}$ 代表的是路的编号, $u_{w3}^{j,k}$ 代表栏目编号.传输效率可达到 97.34%.

在实验中,映射是由群、路标号、栏目代码共计 21bit.

、 、 三种映射的广义几何描述如图 3 所示.

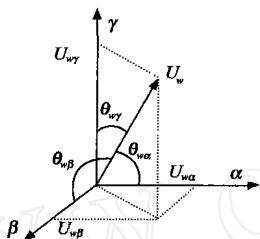


图 3 UCL 映射的几何解释

w 一般比较大,或者说 U_w 一般较短,或者 U_w 较小.

6 信源端 UCL 的标引与复用流程

在数据广播的发送端,需要对数据信息进行采集、标引、复用、格式转换和调制等工作,其工作流程由图 4 所示.

数据广播源端可以粗略地分为标引、一级复用、二级复用、数据格式转换(打包)和调制等阶段.当然,除此以外还有数字媒体服务、授权管理等环节.

按照图 4 所示的流程,我们在 DVB-C 网络环境下进行了传输与解析实验,验证了前述的 UCL 标引方法、两级复用模型和 、 、 三种映射方法的可行性,被标引的 UCL 信息能够正确的映射到各层,在信宿端能够被正确的解析和使用.

不难看出,尽管我们的实验仅在 DVB-C 网络进行,如果更换不同的调制器,可以将我们的实验移植到 DVB-S、DVB-T、DMB-T 等网络.

7 结论

本文研究了在数据广播中进行 UCL 与并播技术的问题.提出了在数据广播中建立基于 Web 的 UCL 结构框架;建立了 UCL 的多映射机制;研究了大规模并播技术;建立了两级复用模型;并在 DVB-C 网络进行了传输与解析实验.实验结果表明,本文提出的 UCL 结构框架、多映射机制及两级复用模型在技术是可行的.

本文的实验研究虽然只在 DVB-C 通过,如果使用不同的调制器,本文研究的结论完全适用于 DVB-S、DVB-T、DMB-T 及其他数据广播体系.同样道理,本文只介绍了基于 Web 的研究,实际上,容易将该研究的结论推广到远程教育、远程医疗、数字媒体、网络图书馆等应用领域.

本文的研究成果对于形成具有核心自主知识产权的数据广播技术标准,推进我国的数据广播与数字电视的增值服务都有着十分重要的意义.

在科技部 863 办公室、国家自然科学基金委和中国数据广播联盟的资助和指导下,课题组正在将本研究成果形成数据广播的 UCL 与传输协议标准草案,核心技术已经申报发明专利.

关于 UCL 在用户终端的应用,如解析、快速下载、基于 UCL 的语义代理等问题将另文讨论.

参考文献:

- [1] C E Shannon. A mathematical theory of communication[J]. Bell Syst. Tech. J. 27 379-423 (Part), 623-656 (Part), 1948.
- [2] Tim Berners-Lee. The semantic web[J]. Science American, 2001 (5) : 21 - 24.
- [3] 李幼平. 共享信息的第二类网络[J]. 中国工程科学, 2002, 4 (8) : 8 - 11.
- [4] 高杨,李幼平. UCL 理念及其系统设计[J]. 电视技术, 2001, 22 (224) : 38 - 41.
- [5] 高杨. 互补结构的信息共享系统[D]. 博士学位论文. 北京:北京理工大学研究生院, 2000. 5.
- [6] 马建国,李幼平,等. 国家规模远程教育平台实验研究[J]. 中国远程教育, 2002, 7 (186) : 38 - 40.
- [7] 李幼平,马建国,等. 国家教育平台[J]. 数据广播, 2002, 2 (16) : 1 - 5.
- [8] 马卫东,李幼平. 数据广播传输体系结构研究[J]. 计算机工程与应用, 2001 (24) : 93 - 96.
- [9] ETSI. Digital Video Broadcasting, DVB specification for data broadcasting[S]. EN 301 192 V1. 2. 1 (1999-06).
- [10] ETSI. Digital Video Broadcasting, Implementation guidelines for Data Broadcasting[S]. TR 101 202 V1. 1. 1 (1999-02).
- [11] 马建国. 具有内容标引的信息共享技术[D]. 博士学位论文. 成都:电子科技大学研究生院, 2004. 6.

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副院长,教授,研究方向为信息系统技术,已出版著作两本,发表学术论文四十余篇. Email :mjg. my @263. net

邢玲 女,1978 年出生于四川成都,中国科学技术大学信息科学技术学院硕士生,主要研究方向为智能信息处理.

(下转第 1643 页)

否存在混沌,值得进一步探索与研究.

参考文献:

- [1] 郑继禹,万心平,张厥盛.锁相环原理与应用[M].北京:人民邮电出版社,1984.
- [2] Cuomo K M,Oppenheim A V. Chaotic signal and systems for communications[J]. In:IEEE Proc of Icassp,1993,3:137 - 140.
- [3] Frey D R. Chaotic digital encoding:an approach to secure communications[J]. IEEE Trans on Circuits and Systems,1993,40(10):660 - 666.
- [4] Cuomo K M,Oppenheim A V,Strogatz S H. Synchronization of Lorenz-based chaotic circuits with applications to communications [J]. IEEE Trans on Circuits and Systems,1993,40(10):626 - 633.
- [5] Cuomo K M,Oppenheim A V. Robustness and signal recovery in a synchronized chaotic system[J]. Int J Bifurcation Chaos,1993,3(6):1629 - 1638.
- [6] Pecora L M,Carroll T L. Driving systems with chaotic signals[J]. Physical Rev A,1991,44(4):2374 - 2383.
- [7] Oppenheim AV,etal. Signal processing in the context of chaotic signal [J]. In:IEEE Proc of Icassp,1992,4:117 - 120.
- [8] 谭永明,邓立虎,郑继禹,锁相鉴频器混沌现象的研究[J]. 电子与信息学报,24(9),2001:1251 - 1256.
- [9] Okasohlu A,Akgul T A. linear inverse system approach in the context of chaotic communications[J]. IEEE Trans on Circuits and Systems (I),1997,44(1):75 - 79.
- [10] 谭永明,葛渭高,郑继禹,锯齿形取样鉴相频率合成器混沌现象的研究[J]. 通信学报,2001,3:20 - 26.
- [11] Mees A,Sparrow C. Some tools for analyzing chaos[J]. Proc Of the IEEE,1987,75(8):1058 - 1070.
- [12] Salam F M A,Sastry S S. Dyna of the forces josephson hynction circuit: the regions of chaos[J]. IEEE Trans. On circuits and Systems,1985,32(8):784 - 796.

作者简介:



谭永明 女,1957年出生于湖南江永县,硕士,副教授,研究方向为通信技术、锁相技术、混沌理论及其在通信中的应用.

邓立虎 男,1955年出生于湖南永兴县,博士,教授,研究方向为微分方程理论及其应用.

郑继禹 男,1937年出生于安徽祁门县,教授,研究方向为通信技术、锁相技术、混沌理论及其在通信中的应用.

(上接第 1624 页)

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和大规模广播技术概念,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.

李在铭 男,1939年出身于重庆市,通信与信息工程学科教授,博士生导师,主要研究信号检测、图像信息识别与压缩技术;多媒体通信,网络与信息综合服务理论与技术.