面向家庭的个性化数字出版系统①

杨 冰[1]，金更达[2]，许端清[1]，傅 强[2]

[1] (浙江大学计算机学院，杭州 310027)
[2] (浙江大学出版社，杭州 310027)

摘 要： 近年来，数字出版在信息社会中的比重越来越大，读者对个性化服务的需求越来越强烈，而目前传统文化出版服务无法满足这种需求，尤其是面向数字家庭的需求。针对面向家庭的个性化数字出版需求，首先采用多维度索引的方法对元数据进行组织表达，将数据挖掘、知识分类管理、自然语言动态关联作为三个维度来建立数据库；采用机械匹配法对数据进行个性化检索，并可以进行主动推送；通过对出版内容进行重组和转码的方法，生成可以保持原版面格式的 EPUB 格式文档；采用数字水印与数字加密相结合的方法来保护作者的合法权益。在此基础上开发了面向家庭的个性化数字出版服务系统，在实践中验证了文中技术的实用性。

关键词： 个性化；数字出版；多维度索引；版面重构；版权保护

Family-Oriented Personalized Digital Publishing System

YANG Bing[1], JIN Geng-Da[2], XU Duan-Qing[1], FU Qiang[2]

[1] (College of Computer Science, Zhejiang University, Hangzhou 310027, China)
[2] (Publication House of Zhejiang University, Hangzhou 310027, China)

Abstract: In recent years, digital publishing share in the information society more and more, readers of the growing demand for personalized service strong traditional culture, publishing services are currently unable to meet this demand, especially for the digital home needs. For family-oriented, personalized digital publishing needs, this paper uses the method of multi-dimensional indexing metadata tissue, data mining, knowledge classification management, natural language as the three dimensions of the dynamic association to create a database; mechanical matching personalized method of data retrieval, and can take the initiative to push; through the reorganization of the publishing and content transcoding method, the layout can keep the original format of the EPUB format document; using digital watermarking and digital encryption to protect the method of combining the author's legitimate interests. In this article based on the development of family-oriented personal digital publishing services system, in practice, demonstrate the practicality of this technology.

Keywords: personal; digital publication; multi-dimensional index; layout refactor; copyright protection

1 引言

近年来，在国家新闻出版总署和整个数字出版行业的大力推动下，我国数字出版产业呈现蓬勃发展之势，不仅产业规模迅速增长，而且在信息社会中的比重也逐渐加大，已成为我国新闻出版产业和文化传播产业的重要组成部分。而随着电信网、广播电视网和计算机通信网的相互渗透与相互兼容，以及数字家庭的逐渐普及，使整个数字出版产业又面临新的发展机遇，并将逐步推动面向家庭的、以个性化消费为核心的数字出版服务方式的形成[1]。然而，这一数字出版服务方式的最终形成，需要着重解决数字出版内容难以进入家庭并满足家庭用户的个性化需求的难题，即

① 基金项目:浙江省社会发展重大科技项目(2007C13051)
收稿时间:2010-04-21;收到修改稿时间:2010-06-05

通过构建面向家庭的个性化数字出版服务系统来打通数字出版内容从内容提供、集成、分发到家庭消费的产业链中的各个环节[2]，满足从数字出版内容乃至传输网络与家庭终端的个性化消费需求。基于此，本文提出了面向家庭的个性化数字出版系统的研究与实现。

面向家庭的个性化数字出版服务系统，即为面向家庭数字设备，譬如个人电脑、PDA 等多种终端，向用户提供关于个性化数字出版的服务以及个性化数字产品的一种系统，其关键技术主要涉及到数据挖掘、个性化检索、格式转换以及版权管理方面。主要解决多种多样、海量数据的存储于组织，以及在此基础上，如何解决用户的个性化需求问题。

2 面向家庭的个性化数字出版服务功能模块

面向家庭的个性化数字出版服务系统按照功能可分为两大系统：数字出版服务系统和数字出版内容管理系统。数字出版信息服务系统完成数字出版门户网站、WAP 网站、电子支付系统、数字版权管理系统和内容分发系统的研发。数字出版内容管理系统的大致功能包括：元数据管理、资源库管理、产品库管理、特色数据库管理全文检索、数据转码、主题定制、书籍快读、内容重构和 OAI 服务等接口的实现。

整个系统负责向用户提供个性化数字出版服务以及个性化数字出版产品。其中个性化数字出版服务包括：专题订阅服务、新书快读服务、网络出版服务、手机听书定制服务、按需出版服务；个性化数字出版产品包括：电子图书、学位论文、收藏艺术品图鉴等。个性化数字出版服务，由客户在线提出个性化需求，根据相应需求转向相应个性化服务，由系统完成相应需求。客户亦可通过获取权限而浏览或使用系统所提供的数字出版产品。

3 个性化数字出版服务系统关键技术

为了开发面向家庭的个性化数字出版服务系统，我们在技术层面上做了如下创新和突破：(1)出版内容表达技术上，提出了多维度索引的概念，将数据挖掘、知识分类管理、自然语言动态关联作为三个维度对元数据进行组织表达；(2)在线内容出版技术上，充分支持个性需求，避免了服务器端指定专题服务；对出版内容采用格式重构和版面重构的办法，生成可以保持原版面格式的通用电子文档，另外对图像、音频、视频进行转码，成为适合网络和电视播出的多媒体格式；(3)通过采用数字水印与数字加密相结合的方法来保护作者的合法权益。本文中研究的个性化数字出版技术详细说明如下：

3.1 基于多维度索引的出版内容组织与表达

所谓多维度，即为多角度。现有的出版技术大多从单一角度表达内容，例如根据关键词，根据神经网络学习的内容等。本文尝试从多个维度，多个角度进行出版内容的组织与表达。具体来说，一方面，将出版内容看作知识，按照 Novins 和 Armstrong[3]的收益最优知识管理模式.Novins 和 Armstrong 的收益最优知识管理模式认为：在知识管理上平均用力，四面开花，往往收效甚微。毕竟，系统的资源是有限的，我们希望系统提供的服务，在准确率的基础上，有着较快的查询速度。因此，针对不同类型的知识采取不同的管理办法，进行强有力的专门领导，是知识管理能够取得最大回报的关键；另一方面，既然出版内容都是自然语言，则可以根据自然语言的动态关联性进行建模：关联规则挖掘算法采用 Apriori 算法。Apriori 算法是关联规则挖掘算法的核心，实际上解决两个问题：①找到所有支持度大于最小支持度的项集，这些项集成为频繁集；②使用①中找到的频繁集产生期望的规则。一旦这些规则被生产，那么只有大于用户给定的最小可信度的规则才被留下。该算法利用层次顺序搜索的循环方法来完成频繁项集的挖掘工作，利用 k 项集来产生 k+1 项集。具体做法为：首先找出频繁 1 项集，即为；然后利用来挖掘，即频繁 2 项集；不断如此循环下去直到无法发现更多的频繁项集为止。每挖掘一层需要扫描整个数据库一遍。

底层数据根据前面所述组织，形成数据库模型，一旦客户端发出请求时，服务器根据主题，结合知识管理模式，运用信息深度智能标引技术，对出版内容进行组织，从而得到更为准确的表达，以牺牲速度的代价得到更高的准确率，充分满足客户的要求。

3.2 个性化内容检索与推送服务

个性化内容检索采用语义分析与知识分类相结合的方法，建立组配检索框，方便用户直观方便的对内容进行检索。用户在使用时，只需点选下拉菜单，输

入较少检索词，系统根据正则表达[4]，自动组配成比较复杂的检索表达式，从而获得更准确的检索结果，达到更高的准确率。此外，可以通过提炼与优化检索词，对检索结果进行二次检索或者分类检索。检索采用机械匹配法，其基本思想是：“先建立一个词库，包含所有可能的词。对于给定的待分词的汉字字符串，按照某种确定的原则对进行分割，得到的子串，若该子串与词库中的某词条相匹配，则该子串是词，继续分割剩余的部分，直到剩余部分为空；否则，该子串不是词，转上重新切取的子串进行匹配”[5]。

在实践中，本文中所研究的内容订阅推送服务技术，首先将用户提供的信息需求，通过多维度索引技术，提炼为精炼的关键词，Web Service 根据关键词定期进行智能检索，检索到合适结果后，由服务器主动通过短信、邮件等手段将用户定制的信息推送到用户所选阅读终端。信息推送有两种方式：①操作式推送 ②触发式推送。与现存的信息运营服务相比，由用户被动接受向定制主动跟踪推送发展，避免用户在获取某一主题过程中普遍存在的重复进行检索、筛选与评价的缺陷。

3.3 个性化内容检索与推送服务

随着掌上电脑、智能手机等家庭阅读设备的普及，以及价格愈来愈低的便利，客户已经偏向于使用家庭终端作为阅读工具。而在当前互联网出版服务中，用户所提交文档一般以 TXT、PDF、DOC、HTML 等通用格式为多，这些阅读格式虽各有其优势，但也都存在各自的缺陷，不利于在家庭阅读设备上阅读，也正由于这些缺陷，导致无法满足数字家庭终端的个性化阅读需求。因此，当前需要解决两个问题：一是如何使越来越多的文档能够在不同家庭的终端上阅读，即跨终端阅读或普适阅读问题；二是如何按个性需求重构文档内容，即基于知识管理的按需出版。本文认为，随着 XML 技术的发展与应用，以国际数字出版论坛(Internet Digital Publishing Forum，IDPF)发布的基于 XML 技术的 EPUB 格式将代表未来数字出版物格式的发展趋势。在本文所涉及的个性化数字出版服务系统中，提出了以 EPUB 格式为核心的支持家庭的个性化需求的重构与转码技术实现框架。

如图 1 所示，系统以基于 XML 技术的 EPUB 格式为基础，或通过 EPUB 文档制作工具直接制作成 EPUB 格式文档，或通过 EPUB 文档格式转码服务将 PDF、HTML、WORD、TXT 等格式文档转化为 EPUB 文档。在面对数字家庭阅读终端，第一，由于 EPUB 格式文档采用 XML 技术，具有较强的普适特性，因而数字家庭阅读终端可以直接阅；第二，可以通过文本语音转码服务，将 EPUB 格式转换为音频格式；第三，根据对内容的个性需求主题，通过内容重构服务，生成符合家庭个性需求的 EPUB 文档。

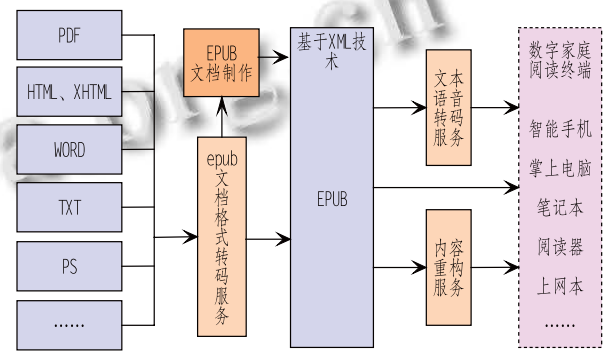

图 1 文档重构与转码技术实现框架

上述功能的实现主要涉及版面分析技术、内容重组技术和格式转码技术：

(1) 版面分析技术。PDF、HTML、WORD、TXT、PS 等格式文档转化为 EPUB 文首先需要进行版面分析，版面分析的算法通常分为自顶向下和自底向上两类[6]。在对版面进行充分分析、识别、理解的基础上就可以进行内容的重构，将其转换成保持原有版面格式的内容可编辑的 EPUB 格式文档。其方法如下：在文本区域，通过搜索连通域，并根据连通域的尺寸特征，优先提取非文本区域，对提取出来的非文本区域，根据投影直方图、宽高比和黑白像素比等特征区分出表格、直线和图像；对文本区域采用改进的机遇投影的纵横切割法来达到对文本正确分割的目的；然后再利用 EPUB 文档格式描述、组织、恢复原有版面的数据和样式；最后生成通过保持原版面格式的 EPUB 文档。

(2) 格式转码技术。格式转码技术涉及文本语音转码技术、多媒体格式转码技术和图像转码技术。其主要目的是调整原有格式以满足目的终端设备使用的要求，能够使视频、图像、音频和文本等媒体内容适用不同的数字设备。

① 文本语音转码技术是将文本内容转化音频文

件，以便具有音频播放功能的数字设备使用，在本文所涉及的面向家庭的个性化数字出版服务系统中，该功能的实现集成了科大讯飞研制的智能语音合成技术。

② 多媒体内容的格式转码将视音频出版物转换为适合网络和电视播出的多媒体格式[7]。在转码过程中，源媒体内容被即时摘要、翻译和转换，甚至去掉某些不支持的媒体对象。

③ 图像转码主要对图像分辨率、格式、色彩等进行转换，使其适合接收终端的显示能力[8]。要将色彩数位 65536、分辨率为 132*768 的 JPG 图像转码成色彩数为 4096、分辨率为 101*80 的 GIF 图像，其通用数学表达式表示为：

$$\mathrm{Im}age_{in}(C1,\ R1,\ F1)->\mathrm{Im}age_{out}(C2,\ R2,\ F2) \quad (1)$$

这里，C1/C2 表示色彩数、R1/R2 表示分辨率，F1/F2 表示图像格式。

3.4 版权保护技术

数字版权管理(digital rights management，DRM)系统允许版权所有者指定终端用户对于数字内容实用的许可、限制以及条件和责任，通过对数字内容存取进行控制，保护作者和内容提供者的知识产权[9-10]。数字版权管理的实现主要考虑文件的使用次数和拷贝权限等问题，这些权限构成数字文件的控制集文件。一般情况下普通用户不具备拷贝权利，这样可以更好地保护数字媒体的版权，对于使用次数或使用权限需要根据用户付款情况确定。针对数字出版物，其数字版权管理由三部分组成：

(1) 将数字内容提供者(如：出版社、作者等)提供的原始电子文档(如：Word 文件、Html、PageMaker 文件等)利用“开发商打包工具”对原始的电子文档进行加密、添加版权信息，如作者、版本号、发行日期等。

(2) 将加密文件部署到内容服务器，加密密钥等存储到另外的授权服务器。

(3) 读者订购后，通过客户端的 DRM 程序向内容服务器提取数字文件，再由授权服务器完成授权，并按照授权文件中的权限使用数字出版物。

本文开发的系统中，所提供数据已经进行加密、插入数字水印，然后和内容标识信息一起打包生成可以发售的数字内容。当用户在满足一定条件后，管理员将会赋予其相关版权许可证，同时将版权许可证与用户身份进行紧密绑定，以防止数字内容的非法共享。用户在客户端点击申请时，输入身份 ID 以及版权许可证编码，方能够顺利获取数字内容。

4 个性化数字出版服务系统系统实现

本文开发的面向家庭的个性化数字出版服务系统，包括个性化在线出版、多终端内容重构、版权保护和在线交易等子系统，支持从出版物出版、在线发行、家庭用户个性化消费整个过程的应用，最终为出版内容提供者、数字出版运营商、网络运营商以及家庭消费者提供全新的数字出版内容服务。数字出版服务平台作为学校网络创新环境的一部分，其总体建设目标为：第一，通过数字出版内容管理系统，实现数字出版资源的收集、制作、管理与维护；第二，通过数字出版信息服务系统实现基于个性化的信息增值服务。系统总体框架如图 2 所示：

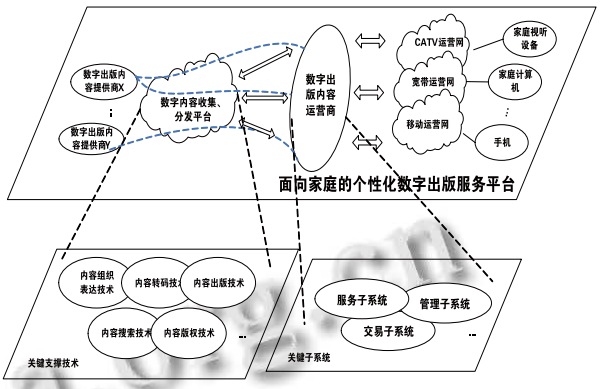

图 2　系统总体轮廓图

系统逻辑架构从上至下由接入层、应用层、服务层和数据层、接入与管理层和安全层构成，以应用层和服务层为核心，以数据层和输入与管理层为依托，以接入层多种方式接入，为各层次的用户提供高品质的个性化的数字出版服务。

输入接口层通过学校网络创新环境学术资源库接口程序从学术资源库中获取数据，并对数据进行审核、分类、入库处理，从而将数据保存在数据层。数据层中由此包括三个部分：元数据库(存储、管理数字出版资源元数据信息和数字出版产品元数据信息；存储、管理用于个性化多维组织的元数据信息)；数字出版资源库(存储、管理学术资源库中科直接用于

数字出版和信息服务的资源；仅用于信息服务的资源)；数字出版产品库(业已存在的产品；根据用户定制后形成的数字出版产品)。服务层负责对元数据、数字出版资源库和数字出版产品库的管理维护，负责对作者授权数字出版证书的管理维护，并包含支持国际数字出版论坛发布的 Open Ebook 格式的制作工具，同时还负责向应用层提供元数据及其内容对象(包括资源库和产品库中存放的内容对象)。应用层包括：数字出版门户网站、数字出版 WAP 网站、内容分发系统、数字版权管理系统、电子支付系统。用户通过数字出版门户网站提出订购请求，触发电子支付系统，完成支付并支付成功后，通过内容分发系统将经过 DRM 加密处理的数字内容下载到客户端，客户端软件向 DRM 许可生成服务器申请许可证书，并下载至客户端。

5 总结

本文研究了面向家庭的个性化数字出版技术，所开发的服务系统充分研究客户的需求，持续跟踪现代数字出版技术发展趋势，研究信息索引，在线内容出版、重构转码等关键技术。在此基础上，设计和实现了面向家庭的个性化数字出版服务系统。本系统的示范应用在浙江大学数字出版服务平台取得很好的效果。示范应用通过与实际客户的个性需求的结合，验证本系统信息服务定位的准确性，提高了经营部门的经济效益。

参考文献

1 焦玉英,索传军.网络环境中信息检索理论与实践的发展.图书情报知识,2001,1:2－6.

2 Fu YJ, et al. Reorganizing Web sites based on user access patterns. ACM, 2001. 583－585.

3 Armstrong R, Novins P. Choosing your spots for knowledge management. Perspectives on Business Innovation, 1998. 45－52.

4 Chen LW, Yuan Q. Person Name Recognition Method Based on Corpus and Rule. Computational Language Research and Development. Beijing: Beijing Institute of Linguistic Press. 1993.

5 郭家义.个性化检索系统中的数据挖掘技术分析.图书情报工作,2003,8: 93－97.

6 刘定强,张忻中.基于组件的中文版面分析.中文信息学报,2000,14(2):8－13.

7 Mostafa ME. MMS. The Modern Wireless Solution for Multimedia Messaging. Proc. 13th IEEE International on Personal Indoor and Media Radio Communication. Lisbon, Portugal. 2002. 2466－2472.

8 张光烈,郑南宁,吴勇,张霞.面向格式转换的数字视频处理方法及其硬件实现.中国工程科学,2001,3(6):41－47.

9 王丽华.基于对等网的数字版权管理方法.现代情报,2008,3:78－83.

10 张建华.网络电子书的数字版权管理技术.科技与出版,2006,2:62－63.