●王红霞　苏新宁

电子政务动态信息采集模型的研究*

摘　要　电子政务动态信息采集模型的设计,要有高效、便利的信息采集平台,支持多种数据格式,提供丰富的检索手段,有支持数字信息的安全管理与审核机制,有开放的多元化的信息资源发布平台,提供资源整合手段。其各项功能由 6 个子模块提供。图 1。参考文献 7。

关键词　电子政务　政务信息　动态信息　采集模型
分类号　G253

ABSTRACT　The authors propose some principles for the design of a gathering system for dynamic information of e-government, including efficient and easy-to-use gathering platforms, multiple-format supports, powerful search functions, security and censorship mechanisms, open and diversified information resource release platforms and resource integration methods, which can be provided in six modules. 1 fig. 7 refs.

KEY WORDS　e-Government.　Government information.　Dynamic information.　Gathering model.
CLASS NUMBER　G253

所谓动态信息资源是指能够连贯地、动态地反映客观事物的性质及运动状态,可以某种方式显示、存贮和传输的,经过人类加工处理的并能够给人类带来某种价值的各种信息的总和。动态信息采集是整个政务信息采集系统的重点组成部分,对系统的动态性能、处理速度、智能化手段等都有较高要求。只有建立动态信息采集模型,通过理论研究和技术手段不断改进系统的动态性能,确保电子政务系统具备畅通的信息网络,才能在面对各种突发性紧急事件时,迅速、及时、准确地将各种相关信息采集到系统中,为建立科学、动态的电子政务决策系统奠定基础。

1　模型的设计思想

1.1　高效、便利的信息采集平台

借助于该平台,可实现在动态采集信息的同时,能够根据信息的主题、形式等特征,使用数据挖掘工具,迅速在系统已有的历史数据、方法库、知识库中找到同新产生的动态信息关系密切的内容,在信息之间建立起一定的历史继承关系,激活静态信息的潜能,实现动态、静态信息资源的一体化管理。通过这些信息的结合,反映出对象的发展变化情况,以及有关事件谁在做、怎样做、效果如何等关联关系,既能给决策者一个整体印象,又有利于建立评价和预测模型。

1.2　支持多种数据格式

电子政务信息采集系统涉及的数据来源丰富,数据量庞大,数据类型众多,数据结构复杂。数据表示形式的多样化,必然导致同类信息分散于不同类型的信息资源中。因此,在设计模型时必须对数据进行整合,解决多种数据格式的识别、兼容与转换问题。解决的关键是构建电子政务元数据体系。在模型中还必须最大限度地确立和使用规范标准,包括支持基于 XML 的数字资源管理。应支持数字资源的标准化表示和传送,数据模型的建立应遵守行业规范并形成规范的流程,支持多种数据格式,包括纯文本、微软的 Word/Excel/PowerPoint, PDF, OpenOffice, JPEG/GIF/PNG 等图像类型、电子邮件、表格、HTML/XML 页面、多媒体文件和各种以附件形式出现的文件。

1.3　提供丰富的检索手段

为了满足用户的个性化需求,最大限度地支持用户发现和使用数字信息,采集过程中需要按照多种分类方式归档。模型还应支持全文检索、智能检索、语义检索、图像音频视频等多媒体检索手段以及二次检索方式,全面支持 XML 格式的数据,提供统一友好的用户界面并把各种政务信息资源整合成统一的结构平台,以实现各种资源的统一检索和各个政务信息资源库之间的信息共享与信息交换,提高信息的查全率、查准率。

1.4　支持数字信息的安全管理与审核机制

系统中机密信息的存在与系统用户的多层次性,使安全问题成为系统构建过程中不可忽视的重要问

* 本文是国家自然科学基金资助项目(编号 70373028)研究成果。

题。动态信息采集必须确保系统的安全、信息来源的可靠以及信息的完整。模型中除了包括防止非法用户侵入、权限控制、存储和传输加密以及数字签名、数字认证等安全防护手段，还应支持知识产权管理，注意版权保护问题，合理合法利用资源。需引入信息的审核机制，鉴别信息来源，在信息采集过程中对不同来源的信息验证其有效性，在信息发布之前确定信息的密级，并根据密级对信息自动加以归类。还可通过群集技术、多机与多数据库的同步备份，合理使用机制等多种方式提高模型的可靠性[1]。

1.5 开放的多元化的信息资源发布平台

为了使用户与系统实现无缝连接、跨平台访问，信息资源发布平台的设计要面向用户，按照系统所采用的目录体系进行信息发布服务。在提供信息服务之前，必须先审核用户的权限，区别用户类型，确定可提供信息内容的范围与密级。提供内容必须符合"新、快、准、全、特、专、精"的要求，满足用户对信息需求的一站式服务。发布形式必须满足用户的不同需求，提供文字、图像、音频、视频等多元化的信息发布形式。内容输出要以方便用户浏览为设计目标，把数据按照用户偏好的风格发布到互联网上。可向订阅某个专题的特定用户或决策人员提供推送服务。

1.6 提供资源整合手段

电子政务数据源分散于政府各级部门、社会各领域，要实现一站式服务，必须整合。通过数字化、网络化、可视化、智能化处理，建立各领域所需的信息资源模式，是电子政务环境下信息资源管理成败的关键，也是电子政务成败的关键[2]。

应在部门间建立通畅的内部信息交换制度，整合统一的电子政务网络，规划或整合政府重要信息基础资源数据库，完善重点业务系统，促进各个业务系统的互联互通、资源共享。对分布在政府内网、外网和因特网上的多源异构数据源进行整合，"多源"包括结构化（各种数据库）、半结构化（文档管理系统，XML）和非结构化（电子邮件、网页、Word等）的信息资源。整合后，通过统一的接口为用户提供服务，用户获取的结果可以同时来源于多个数据库或者是复合数据、非结构化数据[3]。

2 动态信息采集模型的建立

2.1 模型的逻辑结构

从功能定位去考虑，电子政务动态信息采集模型应围绕信息资源库的建设内容进行设计[4]，实现信息资源采集、加工与编辑、审核与发布、管理、服务一整套流程。为满足信息的快速采集、跨平台资源共享以及分布式管理等要求，模型采用先进的三层B/S结构，结合数据挖掘技术、推送技术进行整体架构，能够同时管理文字、图片、多媒体等多种格式的信息，并提供全文检索、智能检索、多媒体检索服务，支持网页的动态发布。它是一个面向内容管理和信息发布的集成化工具，具有资源数字化、存取网络化和管理分布化的特点。其逻辑体系结构如图1所示。

2.2 主要功能模块

模型所设计的各项功能主要由6个子模块提供，它们既各自独立，又相互依存，组合成有机整体。每个子模块又由若干二级模块组成，均能与互联网有效链接。整个模型按分布式数据库原则构建。

信息资源采集模块：实现网上信息采集、现有电子文档格式转换、元数据体系转换等功能。该模块是整个体系的核心，具体包括：电子信息制作工具实现传统信息的数字化；通用文档转换工具可以将现有各种格式的文件转换成统一格式的电子文件；信息获取引擎实现网上信息的采集；全文检索服务器提供TB级海量数据的快速检索服务；流媒体制作平台实现流媒体制作与编辑。

信息资源加工与编辑模块：通过分布式协同在线方式，完成对传统资源的数字化加工、标引分类工作，并支持多种数字资源的融合。

信息资源审核模块：对信息的来源、真实性、有效性、完整性等要素进行审核，确定信息的密级，并可支持在线的发布申请、审核、批准等。

信息资源发布模块：实现对信息资源的各种检索服务，包括全文检索、异构数据库统一检索、分布式检索和基于XML的快速全文信息检索，实现订阅推送、内容发布、全文传送、光盘出版等功能。

信息资源管理模块：实现用户权限管理、数据库维护、统计和计费等功能。

信息资源服务模块：包括内容输出、电子邮件列表、短信传真、信息推送、音频视频实时服务等。内容输出方便用户以传载体形式浏览数据库内容。信息推送服务为信息资源库实现个性化信息推送提供了工具，用户可以向信息资源库提交"订阅需求"，根据订阅信息过滤信息资源库的数据，将用户最需要的信息在第一时间发送给用户。音频视频实时服务支持音频、视频等多媒体信息的流式播放。

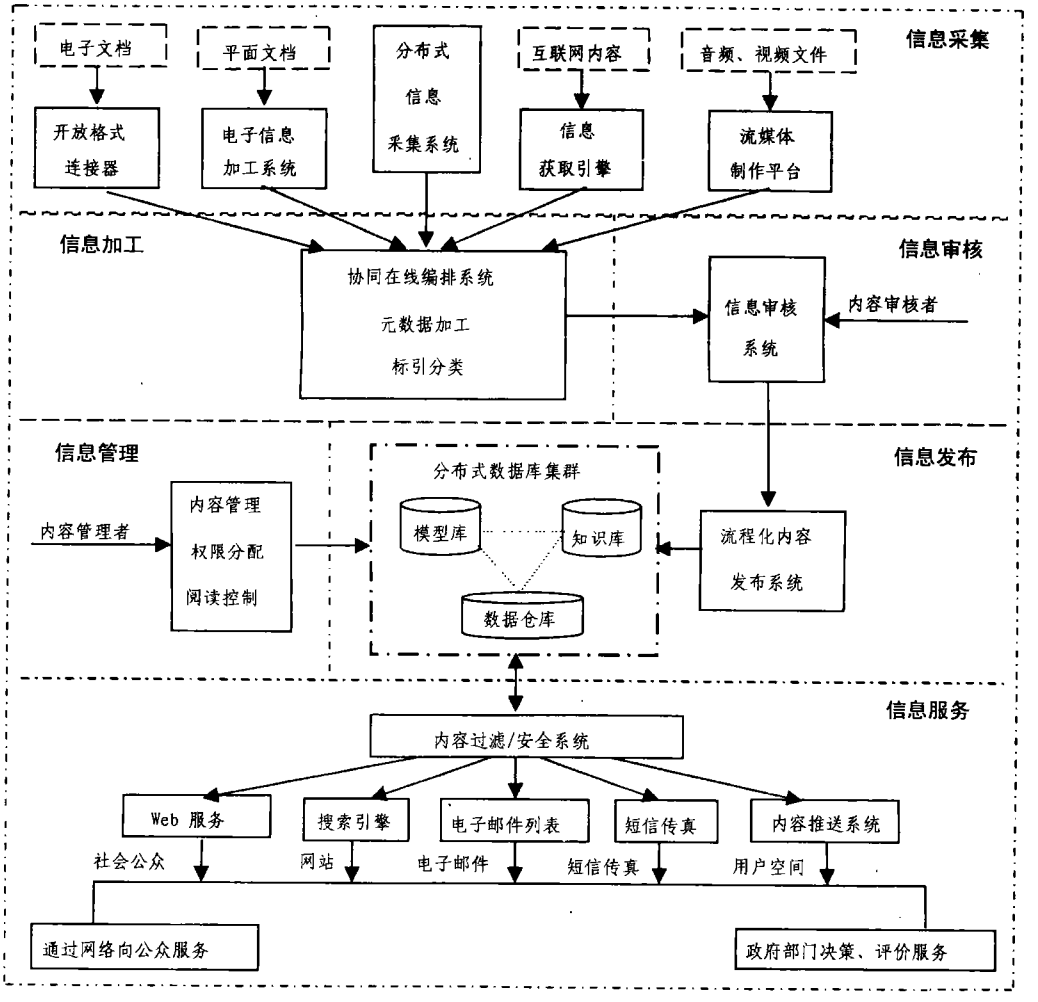

图1　系统逻辑结构示意

2.3　模型采用的访问机制

从整体来看,模型采用三层浏览器/服务器结构(B/S)。B/S结构将模型中的三要素(数据、功能、行为)分离,形成前端客户层,负责可移植的逻辑表达,用户端不用安装任何数据库连接软件,较好地体现了数据处理性能和安全性能;中间的应用层,允许用户通过将其与设计应用隔离而达到共享和控制业务逻辑的目的;后端的设计隔离服务层,提供对专门数据服务的访问,处理客户端与数据库间的数据流。这种先进合理的访问机制,可以很容易地实现跨库检索,为基于该模型构建的系统提供了良好的扩展性。首先,系统可以实现分布式存储,扩大数据的存储量;其次,系统可以方便地利用群集技术实现性能扩展。通过这种方式,可以有效解决超海量数据的存储与大访问量的问题。B/S层次分离的优势还体现在界面风格统一,并具有统一的语言格式、统一的传输协议,系统管理简单,能够优化资源,方便信息发布。

模型中提出的分布式检索集群服务不仅支持同一个服务器的跨库检索和异种数据库检索,而且还支持跨服务器、跨平台的分布式检索,一次允许对网上多个服务器同时检索,更大范围地实现了资源共享。

3　系统实现过程中的关键问题

3.1　多媒体信息的处理

多媒体信息与结构化数据相比,不仅数据量庞大,而且格式多样且有较强的时间性,这些差别导致它在存储结构和存取方式上与结构化数据有很大不同。因此,在系统的实现过程中,必须解决压缩与融

合问题。多媒体信息处理技术是一种文字、音频、视频、图形以及图像与视频等有机结合的技术，更接近于人对信息的处理过程，使人们思想的表述不再局限于顺序的、单调的、狭窄的范围，使信息空间走向多维化，是比应用单一信息效率高得多的信息处理技术[5]。

在对多媒体信息进行处理前，关键是要通过面向对象的方法为多媒体信息建模，即把多媒体信息中的各种复杂关系以形式化的方法表示出来。另外，由于多媒体信息数据量巨大，必须要经过压缩处理才能进行存储和实际应用，目前常用的压缩编码标准有JPEG，MPEG，SC－29，CD－I，RTV等。而对多媒体信息的查询，系统应提供基于内容的检索，实现的关键是为多媒体信息内容建立有效的索引结构。

3.2 分布式人工智能技术的应用

分布式人工智能技术的应用为异种分布式信息资源集成开辟了智能化途径。在电子政务动态信息采集系统中，分布式资源处于网络的不同节点上。为了采用分布式智能技术访问这些分布式资源，需要基于移动智能体构建一个三层嵌套的客户请求、智能体逻辑、应用服务的结构。请求层是用户与系统的交互层，智能体逻辑层是系统业务逻辑中心与移动智能体的管理中心，应用服务层包括系统所需的多种服务资源与应用。为了支持移动智能体在分布计算中作用的发挥，该结构中的每一层都可扩展成一个客户—智能体—服务的计算模型[6]。

3.3 元数据体系的构建

基于数字信息和网络服务的元数据标准，在保证相当检索精度与准确度的前提下，可以方便快速地建立对浩如烟海的数字信息的描述。利用元数据作为电子政务信息资源的组织模式，能够较好地解决信息资源的发现、控制和管理等问题，使政府信息便于内容表达、数据挖掘和知识发现，有利于政府快速有效地决策。

目前，元数据主要应用在图书馆和各种行业的数据组织中，在电子政务领域还没有形成比较规范的体系。元数据体系是进行数字资源管理和网络导航的前提，在数据挖掘、信息检索和信息组织方面有着重要作用，电子政务元数据体系的制订是提供合理著录规则以描述、搜索并处理政务信息资源的核心问题[7]。在系统实现时，可通过XML对元数据进行标识，各种资源元数据的设计基于都柏林核心，但可根据表达内容的不同增加对元数据的设置，以使元数据体系更趋完善。在进行元数据描述的同时，也完成了信息的组织和管理。

3.4 数字信息的存储

无论是对动态的还是常规的电子政务信息，首先应当解决的是对采集来的或自己产生的信息如何组织的问题，如多媒体信息如何组织，数字化的历史原件如何保存，多媒体信息、数字化原件和文本信息如何关联。其次是研究如何将采集来的政务信息科学地、分布式地存放。它们数量大，种类多，分类复杂，集中存放既不可能也不现实。必须将它们分布存放，但这种分布要求科学合理，方便查询和利用。由于政府机构、部门的纵横关系，信息的关联是跨层次、跨行业、跨部门的，所以在分布式存放的实施中，要防止冗余（多处存放）遗漏，要建立分布存放的信息之间的有机联系。第三是数字信息的组织结构问题。如何使用户方便、快速、全面地获取政务信息，不仅仅是检索界面和检索算法的问题，还包括对信息的索引组织。也就是，如何为分散在不同物理位置、不同数据库中的信息构造索引，另外还要解决动态信息的索引及时更新问题。第四是建立专门用于支持政府决策的数据仓库问题。为了公众方便地找到所需要的信息，为公务员快速准确地得到解决问题的知识，还需要研究建立政务信息资源知识地图的问题，通过知识地图把政府的信息链有机联系在一起。

参考文献

1 李湘江. 网络安全技术与管理. 现代图书馆技术，2002（2）
2 蔡运娟，高天鹏. 基于电子政务的政府信息资源管理. 江西行政学院学报，2004（6）
3 王爱云. 电子政务信息资源建设与我国信息化进程. 理论学习，2003（11）
4 张启祥，李慧. 政府信息资源库：电子政务的基石. 信息化建设，2002（5）
5 朱青. 电子政务与多媒体技术. 测控技术，2004（5）
6 操龙兵，南敬昌，戴汝为. 基于移动智能体的分布式资源的智能集成系统. 仿真学报，2002（11）
7 张晓林. 元数据库研究与应用. 北京：北京图书馆出版社，2002

王红霞 南京大学信息管理系在读博士。南京审计学院商学院工作。通信地址：南京大学。邮编210093。

苏新宁 南京大学信息管理系教授。通信地址同上。

（来稿时间：2005－10－28）