电子政务系统数据交换和数据挖掘的模型研究*)

石志国　王志良　薛为民
(北京科技大学信息工程学院　北京100083)

摘　要　电子政务系统在我国才刚刚起步，我们必须开发新的技术来完善电子政务系统的安全技术。本文从两方面讨论了电子政务系统：1. 用 cngXML 建立标准交换条件；2. 借助数据挖掘技术作出明智的电子政务系统决策。该方法已用于北京市某些行政管理区的电子政务系统中。

关键词　电子政务系统，cngXML，数据交换，数据挖掘

Research of the Model of Data Exchange and Data Mining of the E-Government System

SHI Zhi-Guo　WANG Zhi-Liang　XUE Wei-Min
(College of Information Engineering, Beijing University of Science and Technology, Beijing100083)

Abstract　In the beginning of E-government system, to perfect the technology of the security of E-government system, we need new technology. This paper will discuss the E-government system from two aspects: 1. Build the standard exchange document by cngXML; 2. Make E-government wise decision by the technology of the Data Mining. The methods have been used in some districts of Beijing E-government system.

Keywords　E-government, cngXML, Data exchange, Data mining

1　引言

在新世纪里，计算机网络和信息技术越来越普遍，整个世界的社会生活模式和商业模式发生了重要的转变。传统的政府管理手段无法适应目前的变化趋势，传统的服务方式也无法满足企业和个人不断提升的期望。在这个信息化社会里，只有充分认识到这种变化并积极转变思想，政府部门才有可能跟上社会的进步，推动社会的发展。建立电子政务系统迫在眉急。

广义上说，电子政务系统具有如图1的结构。

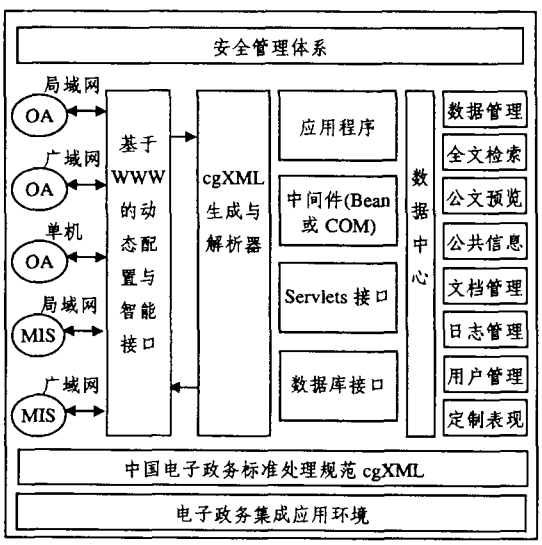

图1　电子政务系统的基本结构

一个高性能、高可靠性、开放的电子政务信息交换平台体系，一般具有如下特点。

·具有电子政务数据交换标准，包括电子政务信息标准和电子政务信息交换标准。

·具有多层次、多角度的电子政务安全体系结构，包括内网安全、内网间安全、内外网间的安全。

·利用电子政务信息挖掘技术，构建电子政务资源综合利用辅助决策支持系统。

本文主要从文档交换标准和数据挖掘模型的角度来分析电子政务系统的解决方案。

2　建立信息交换标准

2.1　电子政务规范语言 cngXML

电子政务需要信息技术作为基础。从政府信息发布、政府网上服务到政府部门间及政府部门内的信息共享和网络办公，都需要不断发展的信息技术作为保障。

与电子政务相关的信息技术包括：Web 应用技术、数据库技术、全文信息检索技术、数据仓库和数据挖掘分析技术、网络连接技术、数据通信技术、安全技术等。在这些技术中，对数据和信息进行灵活、有效、多元化的管理尤为重要，电子政务处理过程中要求能够在异构平台、在不同的网络中实现数据交换和业务自动处理，这些必然涉及到数据、公文和文档格式的标准化、统一化，需要建立一个能够描述政府部门内部、政府部门间和政府与公众间数据交换和业务处理流程的规范标准，以减少数据在处理过程中因标准不统一而引起的诸多问题。将 SGML 强大的表达可选性与 HTML 的简单性进行有机结合的可扩展标记语言 XML，是当前信息技术领域最重要、最活跃的发展之一，已经逐渐成为 Web 上的通用语言。我们的目标就是在 XML 标准的基础上，建立电子政务规范语言 cngXML。

为了建立完善的中国电子政务规范语言 cngXML，中国

*)本文工作得到国家自然科学基金项目(批准号：69975002)和"十五"国家科技攻关项目基金(项目号：2001BA605A)的支持。石志国　博士。王志良　教授，博导。薛为民　博士。

科学院软件研究所电子政务研究室发起中国 cngXML 联盟的倡议，吸引各政府部门的信息管理机构以及从事信息技术研究和应用开发的机构和企业加入，以研究基于 XML 的电子政为交易标准语言——cngXML，建立中国自主的电子政务规范，发展适合于中国国情、具有自主知识产权的电子政务交易规范语言及其支撑平台，推动中国电子政务的发展。

cngXML 的主要部件包括：cngXML 规范、cngXML 词汇表、文档类型定义 DTD（Document Type Definition）、cngXML 中的超级链接、cngXML 文档显示样式、cngXML 文档、cngXML 处理器。

在一般的条件下，电子政务系统的描述和数据交换具有如图2所示的结构就可以满足要求了。

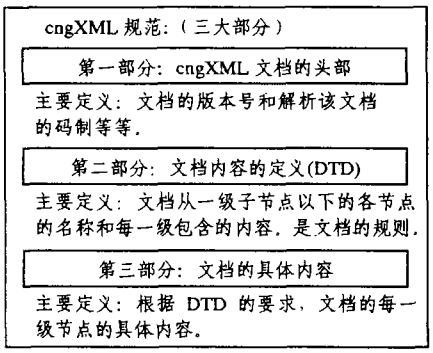

图2　文档交换的格式

虽然 XML 的文档结构非常复杂，但是为了快速高效地进行数据交换，在常规的系统中，采用上面的三大部分作为 cngXML 的标准来进行数据交换就可以了。下面用一个实际的案例来说明。

2.2　电子政务视频会议系统文档交换格式

由于具体的 cngXML 标准还处于研究设计阶段，下面是作者根据上面定义的 cngXML 规范自主设计的某区电子政务视频会议系统的 XML 数据交换文档。

视频会议系统由管理员预先在视频会议服务器上设定，并且可以通过办公自动化系统的会议安排通知与会者，届时与会者向管理员提出申请，由管理员批准其加入会议。管理员可以调度所有会议的参加者，接受或拒绝某人加入会议的申请，强行切断或插入某人的连接。系统可以实现点名发言和举手发言功能，由管理员调度插入发言者的视频和音频信号。其结构如图3所示。

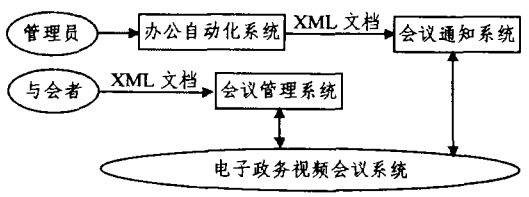

图3　某区电子政务视频会议系统结构图

管理员定义会议通知和与会者申请的数据交换格式统一采用 XML 文档，程序1是笔者根据 cngXML 的规范自行设计的，用来实现会议通知的文档交换。

程序1：数据交换的 XML 文档

```
〈?xml version="1.0" encoding="GB2312"?〉
〈!DOCTYPE 会议[
〈!ELEMENT 会议(主持人姓名,时间,地点,参加会议人,备注)〉
〈!ELEMENT 主持人姓名(#PCDATA)〉
〈!ELEMENT 时间(#PCDATA)〉
〈!ELEMENT 地点(#PCDATA)〉
〈!ELEMENT 参加会议人(#PCDATA)〉
〈!ELEMENT 备注(#PCDATA)〉
]〉
〈会议〉
  〈主持人姓名〉王刚〈/主持人姓名〉
  〈时间〉07/09/2002 15:30AM〈/时间〉
  〈地点〉新办公楼四楼4112会议室〈/地点〉
  〈参加会议人〉区委区政府中层以上干部〈/参加会议人〉
  〈备注〉缺席的人员到区办公室登记。否则通报批评!〈/备注〉
〈/会议〉
```

在发送请求的时候，管理员发送 XML 交换文档到办公系统的解释器。办公系统处理完以后，发布会议通知。其他交换文档格式及其内容可以采用同样的格式和规范编写。

3　电子政务的数据挖掘技术

3.1　数据挖掘技术的电子政务模型

如果将数据挖掘技术引入到电子政务系统中，就需要在 Web 服务器上构建一个数据库系统，用来有针对性地记录政府办公人员和民众浏览和操作的路径。该系统包含多个原始的静态数据库，对于系统管理员给出的一个特定的挖掘任务，需要从中抽取进一层的关联数据库，关联数据库及其操作置于后台数据库系统中。结构如图4所示。

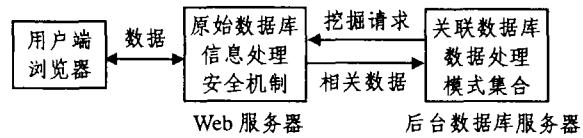

图4　数据挖掘技术方案

将原始数据库与关联数据库分别放置的好处在于：

(1)减轻了网络线路负担。当未给出挖掘任务请求时，Web 服务器上的原始数据库系统自行运行、维护。线路空闲，只有当后台发出任务请求时，才占用线路。

(2)放在 Web 服务器上的原始数据库系统中的信息是公共的，可采用星型结构同时提供给多个用户使用，相互之间不存在影响。

(3)数据挖掘过程及产生的模式均在后台执行，具有安全性。

3.2　系统的数据挖掘过程

对于上面的模型，数据挖掘过程可以分为以下三个步骤实现，如图5所示。

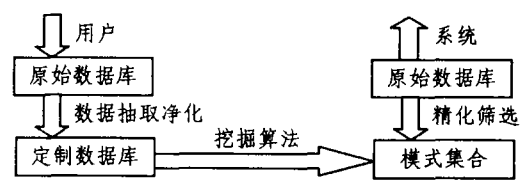

图5　数据挖掘过程

(1)*数据的提取和净化*　这一过程是制定挖掘任务的过程。向 Web 服务器提出请求，从已有的数据库中提取相关数据，可用数据库查询语言 SQL 实现。因为制定挖掘任务时，不同的表达方式可能会造成数据挖掘算法对数据含义理解的不确定性，所以搜集完相关数据后，还需要清除无用数据，对带噪音的数据进行净化。该过程可用统计学的技术检测异常值，

（下转第138页）

还相当有限。除了相机运动等较为简单的特征，用户一般较难提交其所期望的查询。目前的各种运动特征表示方法，在语义上少有明确的意义，这就限制了基于运动特征的视频检索的有效性。从根本上说，视频中的物体识别和分割基本未实现，从而面向对象的运动特征提取也难以实现。基于物体运动的检索现在只是起步阶段，有待进一步的研究与发展。可以展望，今后的运动特征检索将以物体运动为主要的研究目标。

此外，由于各个不同的学者和公司根据自己的需要定义不同的运动特征提取和表示方法，使得很难达到信息共享。2001年9月，MPEG 组织通过了 MPEG-7标准，给多媒体内容的描述者（生产者）和使用者（消费者）提供了一个统一的接口，这其中包括本文前面部分所介绍的各种运动描述子。这在一方面使得运动特征的表示有了很大的通用性，进一步使得基于运动特征检索的广泛应用成为可能；在另一方面，它也使得对运动特征的表达能力有了很大提高。作为一个预测，在将来的发展中还有可能出现形式化的运动特征描述语言，从而使用户对运动特征的表达能力极大丰富，基于运动特征的检索效率也会极大提高。

运动特征检索的另一个可能发展方向是为特定视频片断的检测所服务。比如新闻节目中播音员镜头或体育比赛中的得分镜头的检测。即运动特征将和其它特征结合起来实现复合特征的检索，从而进一步检索具有特定语义的视频片断。单独的运动特征本身往往不具有明确的语义，但许多含有特定语义的视频片断常常具有典型的运动特征。由此可以预见，这种基于运动特征的高级语义检索具有很好的发展前景。

参 考 文 献

1 Chang S F, Chen W, Meng J H. VideoQ: An automated content-based video search system using visual cues. In: Proc ACM Multimedia'97 Conference Proceedings, Seattle USA, 1997. 313～324

2 Rui Y, Huang T S, Mehrotra S. Constructing table-of-content for videos. ACM Multimedia System, 1999, 7(5): 359～368

3 Sudhir G, Lee J C M. Video annotation by motion interpretation using optical flow streams. Journal of Visual Communication and Image Representation, 1996, 7: 354～368

4 Tan Y-P, Saur D D, Kulkami S R, Ramadge P J. Rapid estimation of camera motion from compressed video with application to video annotation, Circuits and Systems for Video Technology. IEEE Transactions on, 2000, 10(1): 133～146

5 Aghbari Z, Kaneko K, Makinouchi A. A Motion-Location Based Indexing Method for Retrieving MPEG Videos. DEXA Workshop, 1998. 102～107

6 Ngo C H, Pong T C, Zhang H J. On Clustering and Retrieval of Video Shots. ACM Multimedia 2001, Ottawa, Canada, 2001

7 Dimitrova N, Golshani F. Motion recovery for video content analysis. ACM Trans. Inform. Syst., 1995, 13(4): 408～439

8 Ardizzone E, Cascia M L, Avanzato A, Bruna A. Video Indexing Using MPEG Motion Compensation Vectors. IEEE International Conference on Multimedia Computing and Systems (ICMCS-99), Florence, Italy, June 1999

9 Milanese R, Deguillaume F, Jacot-Descombes A. Video segmentation and camera motion characterization using compressed data, Multimedia Storage and Archiving Systems Ⅱ. Vol. 3229 of SPIE Proceedings, Dallas TX, Nov. 1997

10 Jinzenji K, Ishibashi S, Kotera H. Algorithm for automatically producing layered sprites by detecting camera movement. In: Proc. of ICIP. 1997. 767～770

11 Jeannin S, Divakaran A. MPEG-7 Visual Motion Descriptors, IEEE Transaction on Circuits and Systems for Video Technology, 2001, 11(6): 720～724

12 Sun X, Divakaran A, Manjunath B S. A Motion Activity Descriptor and Its Extraction in Compressed Domain. IEEE Pacific Rim Conference on Multimedia, 2001. 450～457

13 Yu Tianli, Zhang Yujin. Motion feature extraction for content-based video sequence retrieval. SPIE Proceedings Vol. 4311, San Jose, CA, USA, Jun. 2001. 378～388

14 Zhong D, Chang S-F. Structure Analysis of Sports Video Using Domain Models. In: IEEE Conf. on Multimedia and Exhibition, Tokyo, Japan, Aug. 2001

15 Pilu M. On Using Raw MPEG Motion Vectors To Determine Global Camera Motion, Pilu Digital Media Department HP Laboratories Bristol HPL-97-102 Aug. 1997

16 Ngo C W, Pong T C, Zhang H J, Chin R T. Motion characterization by temporal slice analysis. Computer Vision and Pattern Recognition, 2000, 2: 768～773

17 Lee J W, You S, Neumann U. Large Motion Estimation for Omnidirectional Vision. In: IEEE Workshop on Omnidirectional Vision, June 2000

18 Vieville T, Faugeras O D. Motion analysis with a camera with unknown, and possibly varying intrinsic parameters. In: Fifth Intl. Conf. on Computer Vision, June 1995

19 Wang R, Zhang H-J, Zhang Y. A confidence measure based moving object extraction for compressed domain. In: Proc. IEEE Int. Symp. Circ. and Syst., 2000

20 Wang R R, Hong P, Huang T. Memory-Based Moving Object Extraction for Video Indexing. In: Intl. Conf. on Pattern Recognition (ICPR'00)-Volume 1, Sep. 2000

21 Wei J, Li ZN, Gertner I. A novel motion-based active video indexing method. In: IEEE Int. Conf. on Multimedia Computing System (ICMCS'99), Vol. 2, 1999. 460～465

（上接第97页）

进行平滑处理以及估计噪音参数。

（2）数据挖掘算法　针对不同问题和不同解决方案，上述过程形成了多个由数据挖掘过程中使用到的信息组成的定制数据库，针对这些数据库有很多的数据挖掘算法。几乎所有的数据挖掘算法都要事先提出一些标准来度量产生的模式，通常利用诸如置信度、感兴趣度等的统计属性作为对产生模式的评估标准，从而更进一步决定哪些模式可以保留，哪些模式需要丢弃，更有效地找出潜在的有兴趣的模式。

（3）精化数据和使用结果　对于运行数据挖掘算法生成的过程，可以循环进行新一轮的数据挖掘过程，同时与分析者以及领域专家进行沟通。反复精化、筛选得到认可后，即成为各种形式的知识，其集合就构成了知识库，可以被用户使用。在此模型中就是将有用信息回馈给系统，从而帮助他们准确迅速地决策。

结束语　目前电子政务在全国才刚刚起步，但随着电子政务的发展，全国电子政务工程的推进，对安全性的要求也必将不断地提高。因而，只有不断地发展新的技术，完善电子政务的安全技术规范和相关的法律法规，才能实现安全可靠的电子政务系统。

参 考 文 献

1 Mukherjee B, Hberlein T L, Levitt K N. Network Intrusion Detection[J]. IEEE Network, 1994, 8(3): 26～41

2 Satton G, Buckley C. Term Weighting Approaches in Automatic Text Retrieval . Information Processing and Management, 1988, 24(5): 513～523

3 Freitas A A, Lavinlon S H. Mining Very Large Databases with Parallel Processing[M]. Kluwer Academic Publishers, 1998

4 Kumar A. New Techniques for Data Rerluction in a Database System for Knowledge Discovery Applications [J]. Journala of Intelligent Information Systems, 1998, 10: 31～48

5 HAN J W, Micheline K. Data Mining Concepts a Technique[M]. Beijing. High Education Press, 2001

6 Agrawal R, Imielinski T, Swami A. Database Mining: A performance Perspective [J]. IEEE Trans Knowledge and Data Engineering, 1993, 5: 914～925