

数据挖掘技术在公共管理领域的应用

刘典文

(厦门大学公共事务学院,福建厦门361005)

摘要 数据挖掘技术业已成为提高组织竞争力及组织绩效和服务水平的非常有力的工具。公共管理的诸多领域如电子政府、政府绩效管理和公共危机管理等,同样亟待引入数据挖掘技术,其意义在于:公共管理领域导入数据挖掘技术是政府决策的需要,可以为政府信息共享提供技术支持,是政府信息化的重要保障。在公共管理领域应用数据挖掘技术面临诸多挑战,数据挖掘专业人才的缺乏、政府数据来源及其应用存在的问题、数据挖掘技术自身的问题等。

关键词 数据挖掘;公共管理;可行性

中图分类号:G931.6 文献标志码:A 文章编号:1005-460X(2010)02-0042-05

一、数据挖掘技术概述

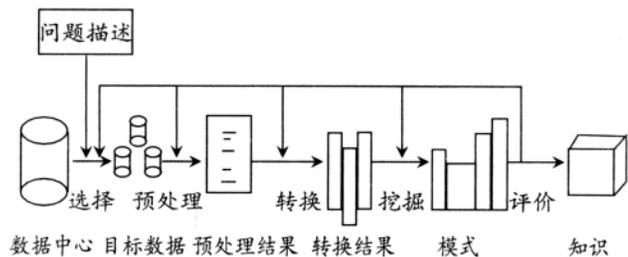
(一)数据挖掘的内涵和过程

数据挖掘(Data Mining,简称DM)是最近几年发展起来的一门新兴技术,其内容涉及人工智能领域中的机器学习、知识发现及数据库领域的数据库、联机分析处理、多维数据库等方面,被认为是解决目前企业信息系统中普遍面临的“数据丰富”而“知识贫乏”问题最有效的手段之一。

1. 数据挖掘内涵。数据挖掘是从数据集中提取隐含在其中人们事先未知的、但又是潜在有用的信息和知识,并将其表示成最终能被人理解的模式的过程^[1]。数据挖掘不但能够从大量不完全、有噪声、模糊、随机的数据中学习已有的知识,而且能够发现未知的新知识,所得到的知识既能为人所理解,又便于存储和应用,这些知识可以用在信息管理、查询响应、决策支持、过程控制等许多领域^[2]。总之,数据挖掘的出发点是代替专家从大量的数据中挖掘出隐含于其中的知识,它使数据存储技术进入了一个更高级的阶段,不仅利用了数据库的存储功能,对历史数据进行查询和辨认,回答“是什么”;而且能够找出历史数据之间的潜在联系,挖掘出其背后隐藏着的许多重要信息,而这些信息是关于数据的整体特征的描述及对发展趋势的预测,在决策

生成过程中具有重要的参考价值,从而可以很好地支持人们的决策,回答“为什么”。

2. 数据挖掘过程。数据挖掘过程是由多个步骤相互连接、反复进行的人机交互过程,一个完整的数据挖掘过程包括问题描述、数据选择、数据预处理、数据转换、运行数据挖掘算法和结果评价等环节^[3]。如下图所示:



数据挖掘过程图

(二)数据挖掘的功能及常用方法

1. 数据挖掘的功能。数据挖掘功能的任务是从大量的数据中发现模式,以帮助管理者获得决策所需的多种知识。在大多数的情况下,用户并不知道数据存在哪些有价值的信息知识,因此,数据挖掘系统的目标应该能够同时搜索发现多种模式的知识,以满足用户的期望和实际需要^[4]。数据挖掘出来的模式可以分为描述型模式和预测型模式,描述型模式是对数据中存在的规则作出一种陈述,或者根据数

收稿日期:2009-07-07

基金项目:国家自然科学基金资助项目《基于数据挖掘的地方政府绩效评估指标设计研究》(10873100)

作者简介:刘典文(1984—),男,江西赣州人,行政管理专业博士研究生,从事政府绩效管理研究。

据间的相似性把数据分组；而预测型模式则能够根据已有的数据集，预测某些未知的数据项的值。数据挖掘的功能模式主要有关联规则、泛化模式、分类模式、聚类模式、时间序列模式和孤立点模式等。

2. 数据挖掘的常用方法。数据挖掘是从数据库的大量数据中抽取潜在的、先前未知的、有价值的知识、模型或规则的非平凡过程，作为一种决策支持过程，主要是基于人工智能、机器学习、统计学等技术，高度自动化地分析政府的数据，作出归纳性的推理，从中挖掘出潜在的模式，预测行为，帮助决策者制定行之有效的战略。数据挖掘常用的方法和技术有几十种，它们分别从不同的角度进行数据挖掘，主要有统计分析方法、决策树方法、神经网络方法、案例推理方法、遗传算法和粗糙集方法等。

(三) 数据挖掘技术应用概述

随着信息技术的迅速发展，数据库技术已经成功地应用于传统的事物数据处理，如行政管理、企业管理、科学、工程数据管理以及许多日益增长的其他应用，这些应用以 GB 甚至 TB 计的速度产生大量数据，如商场的顾客交易数据、证券市场的客户交易数据、Internet 上的巨量信息数据等。正是由于数据和数据库爆炸性的增长，要从这些海量的、以不同形式存储的数据资料中发现有价值的信息或知识，促使数据仓库技术、联机分析处理技术、数据挖掘技术的形成和发展。许多不同应用领域如数据库、知识库、人工智能、机器学习、专家系统、知识获取、统计学、空间数据库、数据可视化等都广泛地应用数据挖掘技术，目前数据挖掘技术已经在农业、医学、天文学、金融、零售、军事、市场营销、Internet 等领域得到广泛应用。随着数据挖掘技术的日益发展和数据挖掘系统的开发运用，数据挖掘的应用领域将会越来越广泛。

二、数据挖掘技术在公共管理领域的应用

(一) 数据挖掘技术应用于公共管理领域的可行性

数据挖掘常用的方法和技术有几十种，它们分别从不同的角度进行数据挖掘，其中有些方法可以应用到公共管理领域，为公共管理决策和行为提供有力帮助和支持。

1. 关联规则。近年来，关联规则的挖掘是学术界研究较多的一个问题，其目的是挖掘出隐藏在数据间的关联或相互关系。如在公共危机管理中，可以通过对政府各种数据库的大量数据进行挖掘，从大量危机记录中发现有趣的关联模式，找出引发危机的关键因素，为预防和避免危机提供参考依据。

2. 孤立点分析。孤立点分析就是从数据集中找出和期望值有显著变化的离群数据，传统的数据挖掘方法都是针对关联规则而设计的，支持度和置信度大，常用于预测，而且这些规则也是专家所能够理解的规则，而孤立点分析得出的意外规则虽然也有较高的置信度，但由于支持度小，常常被用户所忽略，但这些规则往往是出乎专家意料的规则。在一些公共管理事务中，人们更感兴趣的是挖掘出那些意外规则。如对各种欺诈行为进行分析时，通过独立点分析可以发现正常行为与欺诈行为之间的关系，从中得到欺诈行为的一些特征，为管理者进行各种诈骗行为的预警和防范提供决策支持。

3. 聚类分析。典型的聚类分析例证是我国的民族区域自治，即在少数民族聚居的地区实行这种制度，再如公共政策制定也要用到聚类分析，由于公共政策都是针对某一社会问题的，因此，制定政策时必须考虑政策对象的类特征和地方政策的区域特征；再如农作物新品种的推广，如果能够考虑该种子适宜的气温、降雨和土壤的酸碱性等所构成的聚类因素，将提高农作物的推广速度和成功的可能性，降低农业耕作风险。同时，聚类分析对城市规划、公路铁路交通网设计、城市河流系统规划和公交线路设计及站点布局等都是有价值的^[5]。

此外，概念类描述、时间序列分析、数据分类也都可以广泛应用于公共管理领域。

(二) 公共管理中引入数据挖掘的应用领域

1. 电子政务领域。电子政务，是指政府机构运用现代信息和通信技术，打破行政机关的组织界限，改进行政组织，重组公共管理，实现政府办公信息化、政务业务流程信息化，为公众和企业提供广泛、高效和个性化服务的一个过程。其核心内容在于，将政府的管理和服务借助互联网构建一个跨越时间、地点、部门，以顾客满意为导向的政府服务体系。政府部门多年来积累了丰富的“数据资产”，但利用频率、效率却很低，距离决策、预测甚远，没有充分发挥其应有的效用。而数据挖掘技术正是为满足这种需要而产生的一种综合技术，它包含了统计学、机器学习、人工智能、数据库、知识获取、模式识别、分布式多媒体环境的智能代理等。在电子政务中合理使用数据挖掘技术能有效地对政府部门丰富的海量“数据资产”进行开采并加以提炼，使之成为有用的知识，从而对政府的决策起到指导、预测作用。数据挖掘在电子政务中的应用就是把分类、聚类、关联、预测这四类分析技术折射到政府部门，使政府部门的内部信息与外部信息进行有效的整合，从而使政府部门更好地、更有效地将信息发布给最希望得到它

们的公众,使得政府部门更好地服务于公众。

由于政府各部门办公自动化的实现,将产生大量的数据,若对这些数据进行有效的收集和科学的分析,可获得影响政府部门工作的关键因素,为政府部门的决策提供有效的依据,帮助政府部门提高政府信息化的水平,促进整个社会的信息化。数据挖掘被广泛应用于电子政务中的综合查询、经济分析、宏观预测、应急预案、风险分析及预警、质量监督管理和监测、决策支持等系统,它为公众提供一个智能化的、高效的网上政府。例如,几乎每个政府网站都有类似“公众意见调查”的栏目,这对于了解公众需求来说是一个很好的途径,但从网站公布调查结果看,结论大多还停留在对单个问题求总数、求比例等简单分析上。例如有人在分析和研究了关联规则和决策树这两种常用的数据挖掘技术的基础上,将关联规则的 Apriori 算法和决策树的 ID3 算法应用到电子政务数据分析中,建立一个基于数据挖掘的电子政务数据分析系统。利用数据挖掘技术建立一个网上能有效地收集、监测和分析公众大量数据的系统,提炼出实用有效的信息,使电子政务体现人民政府的本色,建成以公众需求为主导的电子政务。将网络数据挖掘技术引入电子政务中,可大大提高整个电子政务系统的智能化水平。

2. 政府绩效管理领域。政府绩效管理是指政府在积极履行公共责任的过程中,在讲求内部管理与外部效应、数量与质量、经济因素与伦理政治因素、刚性规范与柔性机制相统一的基础上,获得公共产出的最大化^[6]。数据挖掘技术在政府绩效评估方面的应用效果非常显著。如果有适当数据和测量标准,是可以对政府项目绩效进行评估的。尤其是当政府组织的服务对象众多或服务项目众多时,利用数据挖掘来评估并进而提升政府绩效是非常有用的^[7]。如,美国教育部通过它的中央处理系统,每年大约处理 1300 万学生的借贷申请。通过使用数据挖掘技术,可以确认人员分布和其他趋势,为领导者修正项目提供信息,从而更好地服务于学生的需求。美国教育部还利用数据挖掘技术来估算教育借贷项目成本和坏账比率(学生到期没有还贷款),这些信息反过来又给分析人员和领导者提供了优化项目的相关信息。美国卫生福利部将数据挖掘引入其项目支持中心,通过挖掘来自雇员资助项目的数据来决定质量标准,提高服务的质量和服务成本效益。

此外,绩效管理中关键绩效指标(KPI)是连接政府目标与员工行为的桥梁,是绩效管理的关键所在。当前在绩效管理中确定 KPI 的方法有:标杆基准法、成功关键分析法、平衡计分卡方法。这些方法对

KPI 的确定起到很大的作用,但是由这些方法确定 KPI 的过程中容易出现一系列的问题:(1)KPI 不能有效反映公务人员的主要工作;(2)KPI 过多,导致 KPI 成了 PI;(3)KPI 不够全面;(4)绩效标准制定过高,缺乏可超越的空间;(5)KPI 的设定不够科学^[8]。究其原因,以上 KPI 的确定方法都是自上向下的进行定性分析,却没有从实际的现实数据上通过定量的统计分析来获得 KPI 的验证。数据挖掘技术则能有效地从大量的实际数据中找到其中的关联性,在数据分析的基础上建立科学有效的 KPI。

同时,一些指标的变化,可能是多个因素综合作用的结果。当关键指标值发生变化时,分析哪些因素的变化导致了关键指标值变化,产生了多大程度的变化,这有利于政府制定解决问题的措施。因此,掌握贡献率非常有利于政府部门进行决策。分析贡献率的技术关键,是建立多个因素对 KPI 影响的函数关系。为解决这一问题,可以采用回归分析法。线性回归是研究一个因变量和一个或多个自变量之间的关系,主要用于分析事物之间的统计关系,侧重考查变量之间的数量变化规律,并通过回归方程的形式进行描述。这样可以生成具体评价对象的评价指标体系,并对已有指标体系、原始指标数据、生成指标数据等综合信息进行统一管理和维护。同时,还可以应用主成分分析和模糊聚类分析方法生成评价对象的评价指标体系,并可以方便地修改评价指标体系的结构;数据挖掘技术还可以用于提取或生成指定评价对象的与指定评价目标对应的指标属性值(形成属性值矩阵)。

3. 公共危机管理领域。在现代社会中,公共危机的发生已经成为社会生活中一种常见现象。按照国际社会的一般看法,公共危机是指社会生活中突然发生的、严重危及社会秩序、给社会造成重大损失的事件。公共危机管理按照时间序列划分为以下五个阶段:危机预警和危机管理准备阶段、识别危机阶段、隔离危机阶段、管理危机阶段、以及处理善后并从危机中获益阶段^[9]。在公共危机管理中,由于危机产生的原因复杂,种类繁多,是一类带有大量不确定因素的半结构化问题或非结构化问题,很难进行科学的计算和评估,很多数据、变量和参数难以量化,传统的分析方法具有很多的局限性,分析和预测的结果也不够准确,迫切需要引进其他技术和方法来解决这类非线性化的问题,而数据挖掘技术为公共危机管理提供了强有力的支持。

危机预警及危机准备阶段。搜集那些可能引发危机的内部信息和外部信息,及时、准确地识别、分析和控制危机,利用数据挖掘技术根据用户的要求

选择相应的数据集,并对数据进行预处理,包括数据净化、离散化处理、泛化处理、空值估算及属性编码等,以建立相应的知识表达系统或决策表,对决策表进行属性约简、值约简和知识约简,抽取最小的属性集合和最小规则集。

在识别危机阶段,主要可以利用数据挖掘技术对第一阶段中提供的危机信息进行加工、整理,在加工整理的过程中要确定信息的真伪虚实,同时使用偏差分析方法从数据集中找出和期望值有显著变化的离群数据。

在管理危机阶段,对于组织高层决策者而言,能够在科学的危机信息调查和准确的危机预测基础上迅速作出决策是非常重要的。网络信息挖掘技术、案例推理方法和决策树方法结合起来给决策分析带来了新的途径,我们可以运用网络信息挖掘评价技术,采取定量分析与定性分析相结合的方法,对网上挖掘出的资源进行再选择与评价,提高决策所需知识的完备性,为科学地进行决策提供基础,可以运用决策树方法分析找出一些具有重要价值的规则信息,帮助决策者制定正确的决策;可以运用基于案例推理的方法在案例库中检索出“最佳案例”,并根据以往的危机管理经验,对当前的危机管理问题提供解决方案和采取的措施,供决策者进行选择借鉴。

在危机后处理阶段,通过运用信息数据挖掘技术,搜索出危机发生原因、危机处理过程,进而对其进行细致的分析,总结经验教训,提出组织在技术、管理、组织机构及运作程序上的改进意见,进而进行必要的组织变革。

三、数据挖掘技术应用于公共管理领域的意义和挑战

(一)公共管理领域导入数据挖掘技术的意义

1. 公共管理领域应用数据挖掘技术是政府决策的需要。政府决策是公共管理中非常重要的一部分,政府管理就是决策,决策失误是政府管理最大的失误。对于政府来说,一次失败的决策不但是经济的损失,而且对政府在人民大众心目中的形象造成重大的负面影响。决策从来都是政府非常关注并投入大量人力、物力和财力去解决的问题。随着政府内部、外部环境的日益严峻及复杂化,影响政府决策的因素越来越多,这些因素之间的关系也越来越复杂。政府要既经济又充分地满足人们对公共服务日益增长的需求,必须深层次地挖掘、分析各级政府积累和采集到的当前和历史的大量各类统计数据,如统计年鉴、台账记录和政府月报等,自动快速地从获取政府决策的有关信息,为科学决策提供快速、准确和

方便的信息支持。随着经济和社会的不断发展进步,以及电子政府的建设和政府各类数据库的建立和完善,政府数据资源日益丰富和庞大,跨政府、跨部门的实时数据存储调用越发常见,政府对系统整合与知识更新的需求越来越强烈,决策所涉及的信息的分析处理远非人力或传统数据分析方法所能胜任,政府必须依靠具有强大数据分析处理能力的技术进行无缝集成,及时处理这些海量的数据资源,以利于决策者作出更加科学、合理、有针对性的决策。

2. 公共管理领域应用数据挖掘技术可以为政府信息共享提供技术支持。随着内部结构的变化及外部环境的不断变迁,政府间的相互协调和政府间的关系日益成为政府管理必须面对和认真解决的问题。而信息共享则是实现政府顺畅协调的重要保证。如果政府之间没有信息共享的理念、机制和相应的信息技术支持,政府间的协调是很难实现的。随着信息技术在政府部门的应用,政府数据的来源具有广泛性和分散性,各种不同来源的数据在格式上存在很大的差别,并产生大量的冗余,难以进行比较、鉴别和利用,同时这些数据来源和存储不规范,不具有统一性,难以从这些不同的数据中发现有意义的结论。在这些因素的影响下,政府虽然建立了内部网,但在政府内部却形成了以部门为单位的一个个数据和信息孤岛,难以实现真正的数据信息共享。如何分享这些分散在异构环境中的数据资源,及时得到准确信息已经成为政府日常管理顺利进行的关键因素,这时就需要借助于新的数据分析处理技术——数据挖掘技术。

3. 公共管理领域应用数据挖掘技术是政府信息化的重要保障。随着信息技术的进步及电子政府的推广和普及,政府信息化已经成为政府提高绩效、降低政府成本的有效途径。信息化政府也是政府建设努力的方向。随着政府信息化观念与实践的推进,已有不少政府业务操作流程实现了自动化,甚至很多管理过程也实现了无纸化和智能化。目前政府的一些领域已从简单的批处理、联机事务处理的信息处理时代迈入了联机分析处理、数据仓库和数据挖掘的信息分析时代。

(二)数据挖掘技术应用于公共管理领域面临的挑战

1. 数据挖掘专业人才缺乏。数据挖掘作为近年来兴起的一种高智能的信息分析和处理技术,目前数据挖掘专业人才相对匮乏。尤其在公共管理领域,虽然许多管理人员掌握了一些诸如描述统计、空白表格程序模型等基本分析技术,但他们在应用如数据挖掘技术等更先进的分析工具时却显得捉襟见

肘,这将极大地限制数据挖掘技术在公共管理中的应用。随着数据挖掘技术的快速进步及其被广泛应用于公共管理领域,对数据挖掘技术专业人才的需求将不断扩大,公共管理领域数据挖掘专业人才缺乏的问题将越来越明显。

2. 政府数据来源及其应用存在的问题。一是政府信息较为分散,没有实现完全的信息共享、数据集成,各级政府及政府各部门的信息系统相互独立,存在数据资源浪费的现象。大量的数据存在于不同地点的数据库中,每进行一次政府数据挖掘都需要收集一次数据,并需要完成大量的数据转化工作,以使数据成为真正有用的信息。面对大量分散的数据资源,各级管理者特别是高层决策者却难以得到及时、准确、综合而实用的信息,这对于提高工作效率、适应瞬息万变的需求、提高公共服务质量是不利的。二是政府众多数据库中大多缺乏可用的、历史的、集成的数据。现有的数据大多是操作型数据,反映的是当前事务的状态。而很多情况下政府数据的挖掘需要大量历史的、存档的、归纳的、计算的数据。而历史数据的管理则较为混乱,由于各级政府及政府各部门软件升级、系统维护等种种原因,导致历史数据分散、不完整、杂乱、格式上的不统一等现象,给现有的政府数据分析预测工作造成了很多的麻烦。三是现在各级政府部门周报、月报、季报、年报数据的信息量非常大,但是反映的仅仅是职能情况等方面的操作型业务数据。政府部门分析预测的影响因素众多,仅仅业务方面的数据不能满足分析预测工作的需要,现在数据的数量巨大、信息量偏少,导致现有的分析预测工作考虑的因素有限,影响了分析预测的准确性和实用性。以上众多因素使得目前应用数据挖掘技术对政府数据分析预测和决策的过程总体水平较低。

3. 数据挖掘技术自身存在的问题。目前数据挖掘技术的研究还不够成熟,其应用还有一些局限性。首先,挖掘的对象方面,数据挖掘要处理的数据量通常十分巨大,成百上千的表,上百万条记录,数据库容量达到若干GB字节,甚至TB字节。更多的属性意味着高维的搜索空间,从而导致组合爆炸,属性之间的关系变得更加复杂,比如表现为层次结构等,这些因素使得搜索知识代价极高。其次,多种形式的输入数据问题。目前数据挖掘工具能处理的数据形式有限,一般可以处理数值型的结构化数据,但大多不能对文本、图形、数学公式、图像或WWW资源等这些半结构、无结构数据进行挖掘操作;再次,用户参

与和相关领域的知识。有效的决策过程往往需要多次交互和多次反复。目前的数据挖掘系统或工具很少能真正做到让用户参与到挖掘过程中。用户的背景知识和指导作用可以加快挖掘的进程,并且保证发现知识的有效性。将相关领域的知识融入数据挖掘系统中是一个重要但没有很好解决的问题。但也正是这些局限性产生了许多生长点,促使数据挖掘研究进一步发展。

数据挖掘作为一种前沿的信息技术,能够从大量的平凡的数据中自动地挖掘出有价值的平时不易为人所察觉的知识,满足了认知的需求,促进了管理的改进,所以在生物医学、工商管理、金融电信等领域取得了卓著的应用成效。公共管理领域引入数据挖掘技术,正如应用其他管理技术和方法一样,不仅是可行的,而且是紧迫的。但就现阶段而言,在公共管理领域应用数据挖掘技术还存在一些不可避免的限制,但与其所带来的益处相比而言,相信在公共管理中应用数据挖掘技术的前景将非常乐观,它不仅改进政府内部的信息化管理水平,对政府向社会提供公共服务的能力和绩效也将起到重要的推进作用。

参考文献:

- [1] FAYYAD MU PIATETSKY-SHAPOMG SMYTH P. Advance in Knowledge Discover and Data Mining [M]. Menlo Park, CA: AAAI Press, The MIT Press, 1996: 1-25.
- [2] 郭萌. 数据挖掘与数据库知识发现: 综述[J]. 模式识别与人工智能, 1998 (3): 292-299.
- [3] HAN Jiawei, KAMBER M. Data Mining Concepts and Techniques [M]. 北京: 高等教育出版社, 2002: 15.
- [4] 贺昌政, 吕建平. 自组织数据挖掘理论与经济系统的复杂性研究[J]. 系统工程理论与实践, 2001 (12): 1-5.
- [5] 胡明晖. 数据挖掘技术及其在公共管理中的应用[J]. 中原工学院学报, 2002 (2): 8-11.
- [6] 卓越. 政府绩效管理导论[M]. 北京: 清华大学出版社, 2006: 1-2.
- [7] 罗伯特·J·卡瓦佐斯, 亨利·约翰·海因兹. 数据挖掘与分析——政府领导力与绩效变革[J]. 朱慧涛, 宫经理, 译. 中国浦东干部学院学报, 2008 (3): 111-115.
- [8] 王怀明. 绩效管理[M]. 济南: 山东人民出版社, 2004: 142-143.
- [9] 薛澜, 张强, 钟开斌. 危机管理——转型期中国面临的挑战[M]. 北京: 清华大学出版社, 2003: 56-96.

(责任编辑:于健慧)