

●毕 强 杨 达 霍红梅

国外信息资源管理的主流研究领域及 热点内容分析

摘 要 介绍了当今国外学者在信息资源管理研究领域取得的主要进展以及热点问题,尤其是国外政府和企业信息资源管理领域的研究进展和研究热点。分析探讨了信息资源管理中的数据组织和数据处理技术。参考文献 9。

关键词 信息资源管理 知识管理 信息处理 信息组织

分类号 G250

ABSTRACT The authors introduce major advances and hot topics in the research of information resource management in foreign countries, especially those on the management of information resources of governments and enterprises, and analyze the data organization and data processing technologies in the management of information resources. 9 refs.

KEY WORDS Management of information resources. Knowledge management. Information processing. Information organization.

CLASS NUMBER G250

信息资源管理是人类对信息管理的一种延伸和发展,是用技术、经济、人文对信息进行管理的产物。本文分析探讨国外在信息资源管理研究领域取得的主要进展,以及研究工作中的热点问题。

1 政府信息资源管理:美国电子政务新战略

随着以因特网为基础的全球信息高速公路建设的全面铺开,政府信息资源管理取得了引人注目的发展。电子政务、电子政务不仅使传统的政府工作面貌焕然一新,而且开辟了信息资源管理的新领域。目前,世界上许多国家在因特网上建立了自己的政府站点体系,实现政府部门内部工作人员之间,以及政府部门与社会公众之间的信息交流。

2002年2月27日,美国白宫管理与预算办公室公布了美国政府新的“电子政务战略”,它按用户不同划分为4大类,根据不同用户有不同的目标:(1)政府—公民:简称G2C,主要目的是建成一站式在线服务,并引入现代关系管理工具以改善提供服务的质量和效率,使公民能得到高质量的政府服务。(2)政府—商界:简称G2B,主要目的是通过大量削减数据收集的冗余度,减轻商界的负担,对商界提供顺畅的一站式支持服务,使用电子商务语言与商界建立数字化通信系统。(3)政府机构之间:简称G2G,主要目的是整合和共享联邦、州和地方三级政府的数据,以改善对信息系统的应用,为关键的政府行为(如救灾行动等)提供更好的综合服务。与此同时,要改善拨款管理能力。该项目还支持“垂直”整合政府业务的需要(如国土安全方面的需

要)。(4)政府内部:简称IEE(内部效率和效能),主要目的是借鉴产业界的先进经验(如供应链管理、财务管理和知识管理),更好地利用现代技术减少政府支出,改善联邦政府机构的行政管理,使各机构能提高工作效率和改进绩效,消除工作拖沓现象,改善雇员的满意度和忠诚度。这些项目计划的总目标是:更有效地使用数十亿美元的联邦政府经费,将政府响应时间从数周减少到数分钟,同时节省目前各级政府业务流程(大部分是纸质文件作业)所耗费的数十亿美元。

2 国外企业的信息资源管理:电子商务与知识管理

企业的目标是利润最大化,这就决定企业信息资源管理有别于政府机构、学术机构等非营利组织,更有特殊性和复杂性。

2.1 电子商务

电子商务是企业信息资源管理的重要组成部分。近年来,国外关于企业电子商务研究的一个突出热点是客户管理技术。

在电子商务环境下,企业的客户就是浏览网站并做出或将要做出购买行为的访问者。从大量的客户行为历史数据中分析、挖掘隐含的有关客户行为的模式并对客户的行为建立模型的技术,我们统称为客户行为理解技术。客户行为理解技术的发展经历了两个不同的阶段:基于Web服务器日志分析的客户行为理解和基于Web挖掘的客户行为

理解。由于 Web 挖掘在目前还处于研究起步阶段,因此,基于 Web 挖掘的客户行为理解技术还很不成熟。但是,从发展趋势来看,它有广阔应用前景。它可以集成各种电子商务数据,充分利用数据挖掘已取得的成果,进行更加完全、深刻的分析,做到真正的客户行为理解。

客户行为理解主要包括以下类型的知识:(1)页面访问次数统计和比较;(2)建立客户个人兴趣信息管理;(3)访问者到购买者的转化率;(4)顾客的回头率;(5)市场策略的投资回报率;(6)个性化商品推荐;(7)个性化广告策略;(8)发现网站服务器隐含的问题;(9)客户个性与访问模式的关联;(10)商品销售情况比较;(11)客户地域分布及人口统计特性与客户购买行为的关系。

在国外,门户技术正在电子商务的各个领域得到广泛应用。企业信息门户(EIP)技术正在受到广泛关注。EIP既可以应用到企业内部,也可以应用到企业之间、企业与客户之间。EIP是一种过渡形式。如同企业信息资源管理正在向知识管理转化一样,EIP在出现之初,就表现出不断向企业知识门户(EKP)转化的趋势。在对数据准确性要求特别高的行业,如建筑业、制造业等领域,EIP和EKP将得到广泛、深入的应用。就目前看来,关于企业信息门户技术的研究热点主要集中在以下几个领域:商务过程、知识过程以及电子商务过程三者间的关系;知识生长周期模型;EIP的结构;EIP的应用及在电子商务解决方案中的地位。

协同电子商务领域的研究也取得了引人瞩目的发展,限于篇幅,这里不作详细介绍。

2.2 知识管理

企业信息资源管理当前的另一个研究热点当属知识管理。

知识管理是信息资源管理发展的新阶段。它与信息资源管理所经历的各个阶段不同,要求把信息与信息、信息与活动、信息与人连接起来,在人际交流的互动过程中,通过信息与知识的共享,运用群体的智慧进行创新,以赢得竞争优势。知识管理是强调以知识为核心的管理,它是知识经济时代的一种全新的管理,既包括将组织中现有的显性知识编码化,也包括发掘员工头脑中的隐性知识使其转化为可编码的显性知识,或者实现隐性知识的共享。

目前商业和企业信息爆炸的程度,意味着知识管理正在成为企业关注的焦点。预计移动和无线等“聪明的”技术激增,包括虚拟小组协作、媒体内容管理、声音门户、地理空间信息管理和个人知识管理、网上会议市场倍增,供应商将提供成套的“聪明系统”,包括内容和文件管理、知识管理和协作产品,许多还包括电子学习。专门技术定位和管理应用将是增长最快的知识管理技术。

知识管理的主要研究领域包括:(1)知识工程和知识管理的语言和结构;(2)知识工程和知识管理的手段和技术;(3)通过机器学习、数据开发和数据库的知识发现获得知

识;(4)知识和知识模型重复使用的方法和技术,特别是与本体论有关的结构、使用和维护的方法和技术;(5)文本的术语、本体论和知识获取;(6)因特网的知识工程和知识管理;(7)内联网知识工程和知识管理的开发利用;(8)组织内不同群体的知识整合;(9)组织内不同群体间的知识共享;(10)知识工程和管理中介方法;(11)知识工程方法在智能信息整合、知识检索和元内容描述中的应用;(12)对知识工程和知识管理方法、技术和工具的评价。

3 信息资源管理中的数据组织技术

3.1 元数据

目前在国外,已有一些较成熟的元数据应用系统。如美国 OCLC 开发的联合在线资源目录系统,它是建立在网络环境下的为本地或网络电子资源创建书目元数据和寻找路径元数据的创建系统,是一种为电子资源编目的数据库和软件包,是建立寻找资源路径的自动工具。北欧国家共同合作开发了都柏林核心元数据与 MARC 记录相互翻译系统。澳大利亚国家图书馆开发了一个逻辑数据模型来识别那些支持存取澳大利亚网络文献资源的特殊实体及其相关元数据。

目前,国外关于元数据的研究集中于讨论 XML 环境下元数据规范的开发与应用。XML 环境下元数据规范包括 WebCollections、CDF、MCF 和 RDF 等。其中 RDF 被认为是最流行、最有前途的元数据数据模型。今后研究工作将主要集中在 RDF 与资源发现技术;RDF 与异构平台上的互操作;RDF 与个性化服务;RDF 与 Web 信息过滤、RDF 与可信 Web、RDF 与智能浏览等方面。

3.2 数据仓库技术

数据仓库是面向主题的、集成的、稳定的、随时间变化的数据集合,用以支持经营管理中的决策过程。最近几年,国外在数据仓库方面开展了大量研究工作,主要包括3个方面。(1)强调利用 OLAP 和 DM 等技术对数据仓库进行分析处理。(2)对复杂数据分析和决策所需要的基本操作进行了考察分析,在相关操作算法与数据结构方面进行研究。(3)研究了数据仓库的体系结构、物理化视图的选择、物理化视图的维护、从数据仓库中的综合数据恢复原始数据、快速收集和有效存取数据等问题,提出了一系列的算法和技术。

Panos Vassiliadis 针对数据仓库的元数据模型只能针对静态元数据和数据间关系建模的缺点,提出了针对复杂动态数据和数据之间关系建模的方法。用 Panos Vassiliadis 提出的方法,可以有效地利用数据仓库进行质量管理,以及数据仓库进行设计、管理和评估。

目前,数据仓库正在向知识仓库演变,即一方面组织处理显性知识,另一方面处理成员头脑中的隐性知识。

3.3 数据可视化

数据可视化可以用于数据挖掘,形成一个可视化数据挖掘方法。在数据挖掘的开始,可视化方法可以通过对可视化数据的观察,提供一个“从哪里开始挖掘”的线索。在数据挖掘过程中,用户可以使用可视化方法对数据进行探索性的分析,直接从可视化数据中发现知识,而不需使用复杂的自动数据挖掘算法。可视化数据挖掘方法可以使用可视化方法表示数据挖掘的结果并验证数据挖掘的方式、方法和结果是否正确。

目前国外研究人员提出了一些数据可视化技术,如几何可视化技术、基于图标的可视化技术、基于像素的可视化技术、分层可视化技术、基于图的可视化技术、用二维和三维图形直接表示数据语义的可视化技术、组合使用上述技术的可视化技术等。数据可视化是一个新的研究领域,刚刚开始不久,还有很多问题需要探索 and 解决。数据可视化是未来几年的重要数据库研究领域之一。

3.4 信息检索与搜索引擎

3.4.1 信息检索

如何组织和提供信息是信息系统要解决的关键问题。目前,信息检索技术可分为3类:全文检索、数据检索和知识检索。知识检索强调的是基于知识的、语义上的匹配,因此在查准率和查全率上有更好的保证。目前知识检索是信息检索研究的重点,特别是面向 Web 信息的信息检索。在对知识检索的研究中,研究人员对本体在知识检索中的作用进行了较为广泛深入的探讨。

3.4.2 搜索引擎

国外网络信息资源检索中搜索引擎的研究进展主要集中在:Web 挖掘与检索的结合;个性化服务;爬行软件的智能化;多媒体和图像搜索;支持跨平台多语言的文档处理;改进与用户的交互;自然语言查询。

国外对于多媒体信息检索的研究相当重视,也获得了不少研究成果。未来几年,多媒体信息检索研究的热点将主要集中在以下几个领域:(1)多途径的多媒体检索;(2)多种多媒体信息库的结合;(3)多特征检索技术的发展;(4)快速算法的探索;(5)多种媒体的检索。

此外,涉及搜索引擎和海量信息挖掘领域的相关技术与方法,如海量网络信息收集、组织与存储,网上文件的搜索与索引服务、主题搜索、网上信息语料库、信息提取、自动文本索引与分类等,也是搜索引擎研究的主要内容。

国外研究人员还比较关注元搜索引擎的研究,认为元搜索引擎将是今后搜索引擎技术发展的方向之一。研究人员将主要注意力放在元搜索引擎的应用研究方面,特别是力图解决由于各搜索引擎语法差异巨大导致元搜索引擎无法进行高级检索的问题。

2003年5月,旧金山的 LookSmart 公司已经开发出新的“分布式”技术,声称可以极大地改善搜索引擎的性能。分布式计算也称为格子计算,在学术界非常流行,旨在针对

复杂的科学等式进行处理。Grub 的奠基人 Kord Campbell 为 LookSmart 公司工作。他声称已经有 1000 名用户下载这个软件。Kord Campbell 表示仍然需要上万名用户来覆盖整个网络,这是传统的搜索引擎只能通过有限的服务器才能完成的。LookSmart 计划引入网络应用程序来分界,网络管理员可以搜索注册过的所有文件。此外,LookSmart 还运作搜索引擎 WiseNut,它采用与 Google 的 PageRank 类似的排名规则,即根据网页相关性、网站内部链接结构及受欢迎程度来确定搜索结果的排列顺序。

美国伊利诺大学香槟分校计算机科学系的助教 Junghoo Cho 博士相信这一模型会有利于搜索。他表示,通过在上千名用户中共享更新后的网络,分布式技术可以针对当前的网络搜索技术的主要问题提出一个大的解决方案。

4 信息资源管理中的数据处理技术:数据挖掘

4.1 数据挖掘

数据挖掘是知识发现中重要的一个环节。数据挖掘是一个交叉学科领域,受多个学科的影响,包括数据库系统、统计学、机器学习、可视化和信息科学等。依赖于所挖掘的数据类型或应用领域,数据挖掘可能集成空间数据分析、图像分析、模式识别、Web 技术、信息检索、心理学等领域的技术。主要的数据挖掘技术与方法有:聚类、分类、决策树、粗糙集、神经网络、回归分析、遗传算法、偏差检测等。从数据库中挖掘的规则可以有多种表达形式:关联规则,特征规则,异常规则,转移规则,序列规则,分类,聚类等。目前,对数据挖掘的研究所面临的难题有:噪音;数据不完整或数据稀疏;超大规模的数据量和高维数据问题;空间数据挖掘;处理不同类型的数据源;可视化数据挖掘。

Sang Jun Lee 与 Keng Siau 讨论了数据挖掘面临的挑战:需要增强处理不同类型数据的能力;研究更有效的算法;提高数据挖掘的结论的价值性;对数据挖掘的结论予以适当表述;在不同抽象水平上进行数据挖掘;自不同数据源进行发掘;保护隐私与数据安全。

目前数据挖掘的研究主要集中在以下方面:(1)挖掘大型的、高维的数据集的算法和系统;(2)挖掘新类型数据的算法和系统;(3)挖掘分布式数据的算法、协议和其他一些基本的架构;(4)简化数据挖掘系统的使用;(5)为数据挖掘建立合适的隐私和安全模型。

数据挖掘领域的发展趋势可概括为:强调实际应用研究,如一般化的、通用的及针对特定领域的数据挖掘系统的开发;提高算法的效率和可伸缩性;改善数据挖掘系统的交互性;提高对复杂数据源和数据类型的处理能力;进一步开发在信息访问和挖掘过程中保护隐私和信息安全的方法;Web 挖掘。

在线分析挖掘 OLAM 是 OLAP 与 DM 相结合的产物, OLAM 兼有 OLAP 多位分析的再现性、灵活性, 以及 DM 的深入性, 是数据挖掘领域研究向深入发展的结果。

数据挖掘在国外已经得到广泛应用。目前的数据挖掘工具主要有 IBM 的 Intelligent Miner, SGI 和 Stanford 大学共同开发的 Mineset, 加拿大 Simon Fraser 大学的 DB Miner 以及 SAS 公司的 Enterprise Miner 等。

4.2 Web 挖掘

Web 挖掘是数据挖掘的一个重要分支。Web 挖掘面对的数据源是异构的、无结构或半结构的因特网数据。这是与传统结构化数据的数据挖掘最明显的不同。

Web 挖掘可分为 3 类: Web 内容挖掘、Web 结构挖掘和 Web 使用记录挖掘。目前 Web 挖掘研究的主要方向包括: Web 文本的自动分类、多层次 Web 信息库的建立、Web log 挖掘, 以及其他涉及信息安全、搜索的效率、查询结果的质量、搜索工具的可伸缩性等方面的问题研究。可以预见, 随着 XML 的兴起, Web 页面会蕴涵更多的结构化和语义信息, 这会使 Web 挖掘变得更有效。

目前在国外, Web 挖掘的研究处于刚起步阶段, 是前沿性的研究领域。将来这一领域引人注目的研究课题有: (1) Web 挖掘中内在机理的研究。(2) 知识库(模式库)的动态维护、更新, 各种知识和模式的融合、提升, 以及知识的评价综合方法。(3) 半结构、非结构化的文本数据、图形图像数据、多媒体数据的高效挖掘算法。(4) Web 挖掘算法在海量数据挖掘时的适应性和时效性。(5) 基于 Web 挖掘的智能搜索引擎的研究。(6) 智能站点服务个性化和性能最优化的研究。(7) 关联规则和序列模式在构造自组织站点的研究。

4.3 分布式挖掘

近几年, 数据挖掘的研究有了很大进展, 对分布式数据库环境中进行数据挖掘的研究显得尤为突出。所谓分布式数据挖掘, 就是使用分布式计算, 从分布式数据库中发现知识的过程。早期对分布式数据挖掘的研究工作主要集中在水平划分的分布式数据库, 其中, 最为突出的发展是多代理技术, 这是提高分布式数据挖掘效率的有效手段, 许多典型的分布式数据挖掘系统都使用了这种软件代理技术。此后, 针对垂直划分的分布式数据库, 研究人员又提出了汇集型数据挖掘系统。如何设计出一个分布式数据挖掘系统的体系结构, 真正能支持各种分布式数据挖掘算法、真正实现平台无关性, 是目前乃至今后的分布式数据挖掘研究的非常重要的一个方面。

5 结论

通过对信息资源管理主流研究领域理论和技术取得的

主要进展及热点研究内容的分析, 我们至少可以得出如下结论: 政府与企业主导的信息资源管理活动, 无论是电子政务还是电子商务、知识管理, 追求的仍都是社会或经济目标, 而绝不是技术目标; 技术手段恰恰是信息资源管理领域发展最快、最具扩张性的因素, 信息资源管理研究正在表现出越来越明显的纯技术倾向; 各种信息资源管理技术都在向具有智能化特点的方向发展, 这是一个共同点。信息资源管理技术还表现出了综合化与语义化的特点, 即同时采用多种技术相互协调解决问题, 或多种技术相互应用或相互结合满足信息资源管理新的要求, 以及面向下一代互联网即语义互联网的技术越来越多地应用到信息资源管理领域。

参考文献

- 1 Mike Uschold. Knowledge level modelling: concepts and terminology. The Knowledge Engineering Review, Vol. 13:1, 1998, 5-29
- 2 William, S. And Austin, T. Ontologies. IEEE Intelligent Systems, 1999 Jan/Feb:18-19
- 3 Uschold, M, King, M, Moralee, S and Zorgios, Y. The enterprise ontology. The Knowledge Engineering Review, 1998, 13(1)
- 4 M. Uschold and A. Tate. Putting ontologies to use. The Knowledge Engineering Review, Vol. 1998, 13:1: 1-3
- 5 Martha E. Williams. Highlights of the Online Database Industry and the Internet:2000. In: Information Today. Proceedings-2000: Proceedings of the 21st National Online Meeting: May 16-18, 2000, New York. Medford, NJ: Information Today Inc, 2000. p. 1-5
- 6 Berners-Lee T, Hendler, J., and Lassila, O. The Semantic Web. Scientific American (284:5), 2001, pp. 34-43
- 7 Madnick, S. The Misguided Silver Bullet: What XML Will and Will Not Do to Help Information Integration. in Proceedings of the Third International Conference on Information Integration and Web-Based Applications and Services, Linz, Austria, 2001, pp. 61-72
- 8 Mihaila, G., Raschid, L., and Vidal, M. E. Source Selection and Ranking in the WebSemantics Architecture Using Quality of Data Metadata. Advances in Computers (55), 2001, pp. 87-118
- 9 Pottinger R., and Halevy, A. Y. MiniCon: A Scalable Algorithm for Answering Queries Using Views. LDB Journal (10), 2001, pp. 182-198

毕强 吉林大学南岭校区管理学院信息管理系教授, 博士生导师。通讯地址: 长春。邮编 130022。

杨达 沈阳中国刑警学院图书馆馆员, 吉林大学在职硕士研究生。通讯地址: 沈阳。邮编 110000。

霍红梅 辽宁省委党校图书馆馆员, 吉林大学在职硕士研究生。通讯地址同上。

(来稿时间: 2003-06-06)