

国内外数据整合研究进展分析

李伟华, 郑彦宁, 刘志辉

(中国科学技术信息研究所, 北京 100038)

摘要: 互联网技术的快速发展使企业和科研工作都迈向了数据化,但同时也带来了数据分散、孤立、整合困难的问题,数据整合一直以来都是情报学的重要研究内容,从情报学的视角发现国内外数据整合研究的差异就显得十分重要。本文利用共词分析的方法对国内外数据整合的研究进行对比分析,结果表明国内相关研究整体落后于国外,在从基础技术向应用过渡的过程中,国内主要侧重于商务智能、面向服务应用等商业方面,而国外则更多地关注于地理信息系统、生物信息学、基因表达等公共事业方面。

关键词: 数据整合; 主题分析; 共现分析

中图分类号: G25

DOI: 10.3772/j.issn.1673-2286.2015.06.009

1 引言

早在1950年前后就有学者开始研究数据整合,并持续得到各学科学者的广泛关注。信息化和互联网化提高了各个行业中企业的工作效率,同时也引发了信息分散、孤立、无法整合利用的问题。数据化时代的到来更是让人们生活和工作时刻离不开数据,并随时都在产生数据,而如何把不同类别、不同形式、不同来源的数据整合起来应用到现实生活中却是一个比较困难的问题。如果能够将数据整合应用于生物、医疗、公共卫生、交通、城市规划等方面将极大地改善人们的生活,而数据整合在竞争情报和商业数据分析中的成功应用一定程度上将提高决策的科学性和效率。各学科学者和商业人士已经发现数据整合产生的巨大价值,但是如何整合现实生活中大量的复杂数据并将其应用于商业和科研决策中依然存在很多困难。信息技术的快速发展给解决数据整合的难题带来了希望,有关数据整合的研究近年来一直处于比较活跃的状态。但截止到现在还没有学者从科学计量学的角度对有关数据整合的研究进行分析。本文将对国内外有关数据整合的研究进行分段并分析不同年段的研究主题,然后将国内

外研究进行对比分析,以期以后有关数据整合的研究提供借鉴。

2 数据来源及分析方法

2.1 数据来源

国外的数据主要来源于Web of Science的核心数据库集,在数据库中以“TI=Data Integrat* OR TI=Data Consolidat*”为检索式,不限年限进行检索,共得到文献9484篇,检索时间为2015年1月15日。国内的数据主要来源于CNKI(中国知网),以“篇名=数据整合 or 关键词=数据整合 or 篇名=数据集成 or 关键词=数据集成”为检索式,不限年限进行检索,共检索到文献8032篇,检索时间同样为2015年1月15日。

2.2 分析方法

(1) 共词分析

共词分析是一种内容分析技术,通过分析在同一文本中词汇对共同出现的形式,来确认文本所代表的学

科领域中相关主题的关系,进而探索学科的发展,通过对高频主题词的聚类可以发现学科的研究热点^[1]。一般认为词汇对在同一篇文献中出现的次数越多,说明这两个词汇的关系越紧密,因此统计一组文献的主题词两两之间在同一篇文献出现的频率,便可构造由这些主题词组成的共词网络,通过网络节点之间的远近便可反映主题内容的亲疏关系。共词分析法就是以此为原理,将文献主题词作为分析对象,利用包容系数、聚类分析等多种统计分析方法,把众多分析对象之间错综复杂的共词网络关系以图形直观的方式表示的过程^[2]。

(2) 演化分析

演化分析是一种通过对比分析事物在不同时间段的异同来反映事物发展变化情况的分析方法,演化分析起源于生物学,现已被广泛的应用到管理、技术创新等方面的研究^[3,4]。

3 数据整合研究进展分析

3.1 数据整合研究的发展趋势分析

对检索到的数据按年份分国外、国内文献以及国内学者国外发文进行数量统计,如图1所示。

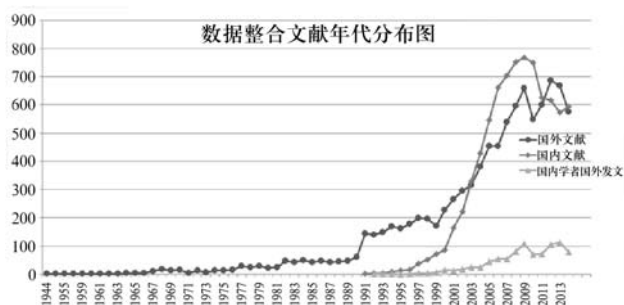


图1 数据整合研究文献年代分布图

从图1可以看出国外有关数据整合的研究最早出现于1944年,关于医学领域数据结论整合^[5],从1959年开始数据整合得到各学科科学家的广泛关注,并于1970年左右相关研究出现缓慢增长。在1991年相关研究数量急剧上升并在此基础上缓慢增长,到2000年出现快速增长,于2009年达到局部高峰。而国内的相关研究首次出现在1987年,从1991年到1999年呈现缓慢增长,于2000年出现快速增长,2009年达到高峰,之后呈现

下降趋势。另外从图中可以发现,国内有关数据整合的研究起步比国外晚40年左右,但是在2000年以后相关研究文献的增长速度比国外还快,并在2003年国内相关研究文献数量首超国外。然而从国内学者在国外期刊发文数量变化的曲线和国外、国内文献两条曲线的对比发现,其变化趋势和国外的研究曲线更相似而不是国内。而从文献增长率也可以看出,虽然国内高于国外,但国内学者的国外发文增长率却明显低于国外相关研究文献的增长速度。由此可见,国内学者十分关注数据整合的相关研究,但是高质量、高价值的研究却比较少。

3.2 数据整合在不同研究阶段的主题分析

结合图1中国内外文献两条曲线的走势,本文将把有关数据整合的研究按年代划分成5段:1990年以前、1991-1999年、2000-2003年、2004-2009年和2010-2014年。在分段的基础上利用共词分析对每个阶段的研究主题进行归纳,并将国内外的相关研究进行对比。另外,在进行数据处理的时候发现从Web of Science中获取的数据中1990年之前的文献关键词字段为空,经查阅得知,该数据库数据1990之前的文献没有关键词字段,而不是下载或处理的问题。所以本文只针对1990年以后的数据进行主题分析和国内外相关研究的对比分析。

每个阶段国内外数据整合文献的数量、选取的TopN关键词以及选取标准如表1所示,对Web of Science和CNKI不同阶段的数据会进行相应处理并绘制高频关键词共现网络图,其中具体的处理过程在3.2.1中作简要介绍,其他各阶段均按此法处理。

表1 各阶段文献数量分布表

时间段	数据源	文献数量	TopN关键词	选择标准
1991-1999年	WOS	1150	146	频次>=2
	CNKI	208	142	频次>=2
2000-2003年	WOS	1106	209	频次>=2
	CNKI	800	279	频次>=3
2004-2009年	WOS	3086	329	频次>=3
	CNKI	3860	516	频次>=5
2010-2014年	WOS	3081	387	频次>=3
	CNKI	3164	600	频次>=3

3.2.1 1990-1999年间数据整合研究主题分析

把从Web of Science中获取的1991-1999年间的1510篇文献中的关键词数据导入汤森路透的TDA (Thomson Data Analyzer) 工具中进行清洗和统计, 并选择出现频次大于等于2的关键词构建Top146*Top146的矩阵, 并结合Ucinet和Netdraw两个工具绘制高频关键词共现网络图, 如图2所示。由图中可知此阶段有关数据整合的研究主题还比较少, 彼此之间的联系也不是很紧密, 其中学者关注较多的是数据整合和数据库 (Data Base)、数据交换 (Data Exchange)、数据模型 (Data Model) 之间的关系, 而数据整合的应用则主要集中于地理信息整合 (GIS)、遥感信息整合 (Remote Sensing)、计算机辅助设计 (CAD)、计算机辅助制造 (CAM)、声音数据整合 (Voice Data Integration、Integration of Voice and Data、Voice and Data Protocol)、信息系统 (Information Systems) 和数据流程 (Data Processing) 等方面。从图中也可知, 地理信息系统和遥感数据之间关系紧密, 计算机辅助设计和计算机辅助制造往往共同出现, 也有学者关注机器学习 (Machine Learning) 和系统整合 (System Integration)、数据安全 (Data Security) 和异步传输模式 (ATM) 之间的关系。

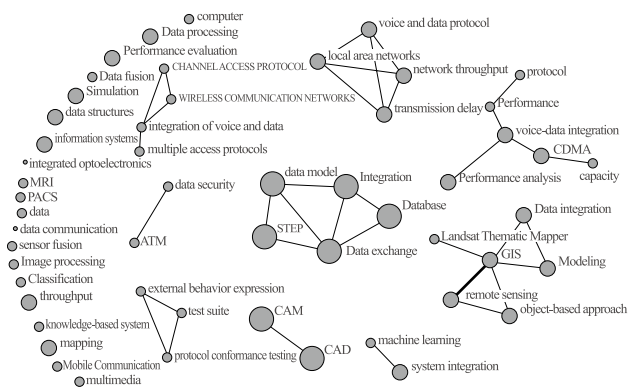


图2 1990-1999年间国外数据整合研究主题网络图

把从CNKI中获取的1991-1999年间有关数据整合的研究文献208篇研究文献的关键词提取出来, 经过清洗和自编程序处理得到出现频次大于等于2的关键词构建Top142*Top142的矩阵, 并结合Ucinet和Netdraw两个工具绘制高频关键词共现网络图, 如图3所示。相比此阶段国外研究各研究主题之间的松散, 国内研究主题稍为紧密, 此阶段国内学者主要关注的和数据整合

相关的主题是数据仓库、联机分析处理、决策支持系统、计算机辅助设计、计算机辅助制造和异构数据源、数据环境、信息系统和地理信息系统等。其中联系最为紧密的是数据仓库、决策支持系统和联机分析处理形成的三角关系, 其次为分别以计算机辅助设计、异构数据源和Voice为核心的研究主题。还有一个比较重要的连接中心是Sybase公司, 该公司和数据库相关。此时间段内国内外有关数据整合的研究并没有太大差异, 但是国内学者对于数据整合的基础数据库技术关注度非常高。

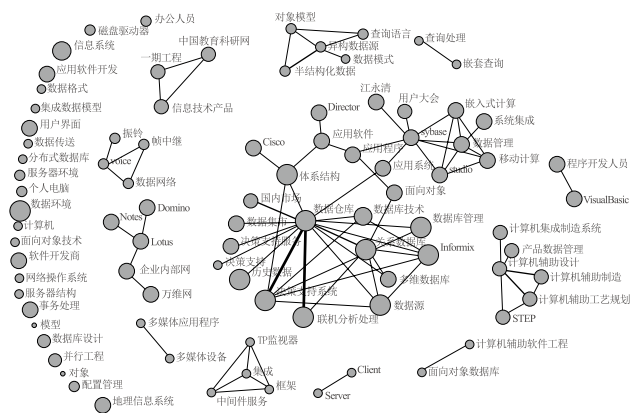


图3 1990-1999年间国内数据整合研究主题网络图

3.2.2 2000-2003年间数据整合研究主题分析

相比上一时间段, 此阶段的研究主题则更为丰富, 从图4可以看出, 国外文献研究数据整合 (Data Integration) 有直接连线的是地理信息系统 (GIS)、数据挖掘 (Data Mining)、可扩展标记语言 (XML) 和企业资源规划 (ERP), 而围绕这些主题的是一些更细的概念, 如遥感信息整合、全球定位系统 (GPS)、数据仓库 (Data Warehouse)、知识发现 (Knowledge Discovery)、联

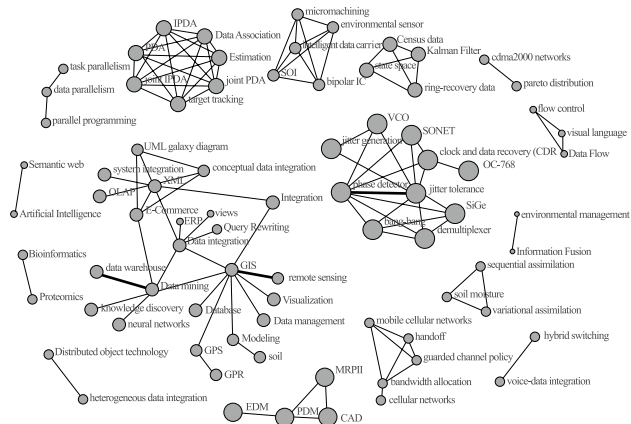


图4 2000-2003年间国内数据整合研究主题网络图

机分析处理 (OLAP)、电子商务 (E-Commerce); 从计算机辅助设计和计算机辅助制造发展而来的是产品数据管理 (PDM) 及相关的一些研究制造资源计划 (MRPII)、计算机辅助设计、电子设计制造 (EDM); 对异步传输模式更加具体的研究是由鉴相器 (Phase Detector)、抖动容限 (Jitter Tolerance)、压控振荡器 (VCO)、同步光纤网络 (SONET); 以及由智能数据载体 (Intelligent Data Carrier)、双极集成电路 (Bipolar IC)、绝缘体上硅 (SOI)、微加工 (Micromachining)、环境传感器 (Environmental Sensor) 组成和以掌上电脑 (PDA、IPDA、Joint IPDA) 为核心的等两个比较新的研究主题群。

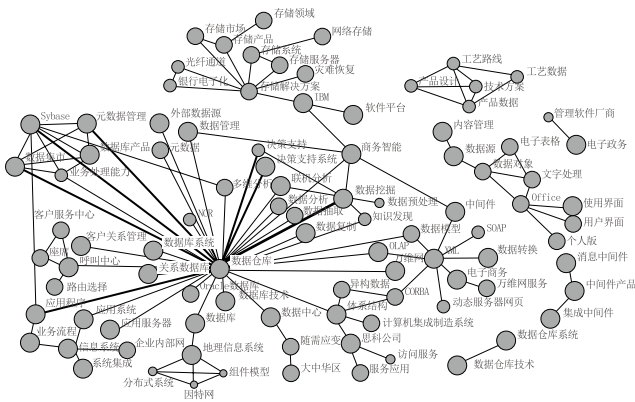


图5 2000-2003年间国内数据整合研究主题网络图

从图5可以看出, 此阶段国内和数据整合最为密切的研究主要集中在和数据存储相关的数据仓库技术 (Sybase公司、数据集市、数据库产品、数据库系统、关系数据库、Oracle数据库等) 和数据存储技术 (存储解决方案、存储服务器、网络存储、存储市场、灾难恢复等) 以及和网络数据整合相关的XML、OLAP、电子商务、万维网服务和动态服务器网页, 还有和Office相关的文字处理、电子表格、内容管理等。此阶段国内数据整合主要应用于地理信息系统 (分布式系统、因特网、组件模型、数据库)、客户关系管理 (客户服务中心、呼叫中心、座席)、产品数据管理 (产品数据、产品设计、工艺数据、工艺路线、技术方案) 以及其他一些面向企业应用的系统研究 (应用系统、信息系统、系统集成、业务流程、企业内部网、应用程序)。而为应用提供技术支持的是以数据挖掘为中心的数据挖掘技术 (知识发现、数据抽取、数据预处理、数据分析等) 以及以数据挖掘为基础的商务智能和决策支持系统, 另外从图中分析可知进行数据挖掘和整合的基础是存储于

数据仓库中的数据。图中还出现了三家国外的知名企业 Sybase、IBM和思科, 其中Sybase和思科出现于前一阶段, 在此阶段处于更加重要的连接位置, 而与IBM有直接关系的则是软件平台、存储解决方案和商务智能。

3.2.3 2004-2009年间数据整合研究主题分析

由于数据整合 (Data Integration) 在网络图中的影响力太大, 导致无法呈现与数据整合相关的研究热点之间的关系, 故在此阶段和下一阶段的主题网络图绘制中没有让Data Integration (国外研究)、数据整合 (国内研究) 节点出现在关键词共现图中。由主题图6可以看出, 此阶段的研究热点比较集中, 并且各主题相互之间联系紧密, 其中核心连接点是本体 (Ontology)、数据挖掘 (Data Mining) 和地理信息系统 (GIS)。围绕本体的是语义相似性 (Semantic Similarity)、语义整合 (Semantic Integration)、语义网络 (Semantic Web)、资源描述框架 (RDF) 等语义研究, 还有就是网络数据整合 (Web Services、XML、OWL)、元数据 (Metadata), 另外我们可以看出本体和生物信息学 (Bioinformatics)、数据挖掘也产生比较直接的联系, 而生物信息学和数据挖掘又通过Database (数据库) 产生间接地关联。和数据挖掘联系最为密切的是聚类 (Clustering)、分类 (Classification)、决策支持系统 (Decision Support System), 其次是机器学习 (Machine Learning)、关联规则 (Association Rule) 和数据仓库 (Data Warehouse), 其中数据仓库又是联机分析处理 (OLAP)、商务智能 (Business Intelligence) 的基础支持, 地理信息系统 (GIS) 和遥感 (Remote Sensing) 关系依然非

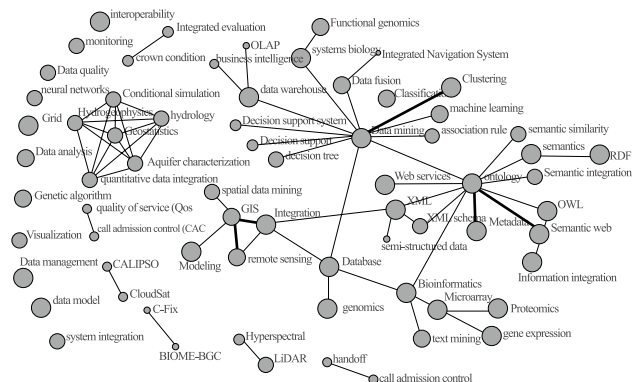


图6 2004-2009年间国外数据整合研究主题网络图

常紧密。还存在着由统计学 (Geostatistics)、水文学 (Hydrology)、水自然物理 (Hydrogeophysics)、条件模拟 (Donditional Simulation)、含水层表征 (Aquifer Characterization)、定量数据整合 (Quantitative Data Integration) 等概念组成的水文、地理数据整合研究。

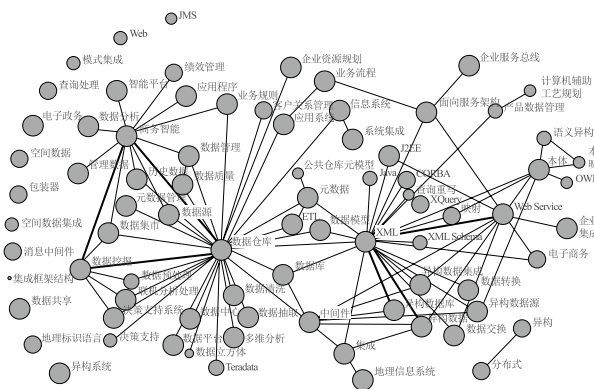


图7 2004-2009年间国内数据整合研究主题网络图

此阶段国内有关数据整合的研究热点也比较突出, 其中比较核心的研究热点是数据仓库、数据挖掘、商务智能、可扩展标记语言 (XML)、网络服务 (Web Service), 如图7所示。其中数据仓库和数据挖掘在上一阶段就是学者比较关注的热点, 而围绕这两大主题的小概念也基本没有太大变化, 然而商务智能在此阶段得到更多学者的关注并且和数据仓库、数据挖掘之间形成明显的三角关系, 也就是说数据挖掘技术应用于由数据仓库中的底层数据带来的商业智能应用。而此阶段处于比较核心位置的另一个节点XML, 表现出与上一阶段截然不同的研究内容, 此阶段和XML联系最为紧密的是Web Service、编程语言、异构数据、元数据等主要概念。其中中间件也是一个从开始就出现, 并经历从出现到演变成成为关注热点的变化, 从图7可以发现中间件和数据仓库、数据库、地理信息集成、异构数据、Web Service、XML等都有直接的联系, 在图中还有两个新出现的处于比较重要连接点位置的点: 本体和面向服务架构。

3.2.4 2010-2014年间数据整合研究主题分析

国外文献有关数据整合的主题网络图如图8所示。从图中可以看出, 此阶段最为核心的连接点是Data Mining (数据挖掘) 和GIS (地理信息系统), 而有关本体的研究热度则有些下降。另外, 有关生物信息学的研究则成为学者的研究热点, 并且与数据库主题紧密联

系, 另外两个联系比较紧密的点就是基因表达 (Gene Expression) 和芯片 (Microarray), 而有关基因表达的研究也属于生物信息学的范畴。有关云计算 (Cloud Computing) 的技术也被应用到数据整合中来进行资源管理、动态数据整合和可视化, 而GPS (全球定位系统) 于2000-2003年段出现后再次成为研究热点, 此阶段有关GPS的研究主要集中在地面变形 (Ground Deformation)、干涉合成孔径雷达 (InSAR) 和差分干涉测量技术 (DInSAR) 等方面。

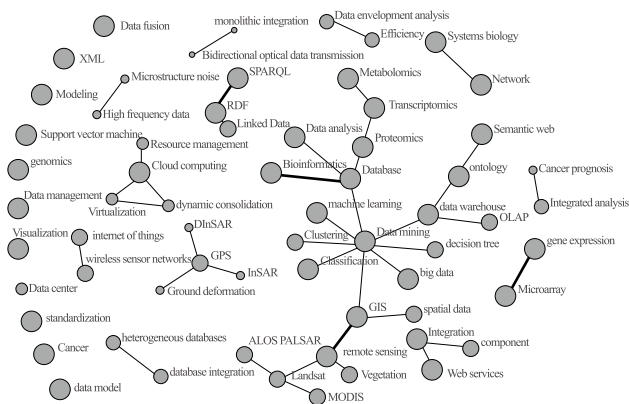


图8 2010-2014年间国外数据整合研究主题网络图

与上一阶段相比, 国内有关数据整合的研究此阶段并没有增加比较新的研究热点, 如图9所示。其中最主要的变化是有关商务智能的研究热度下降, 而以XML和Web Service两个节点为核心的研究主题群反而受到更多学者的关注, 尤其是XML、中间件和异构数据概念三者之间的关系更加紧密。而有关本体和面向服务架构的研究也明显增多, 明显可以看到围绕本体的是有关语义、深度网络的研究, 并且有关本体的研究和元数据、异构数据、中间件、XML之间有着密切的联系; 而面向服务架构则主要是面向企业服务并和Web

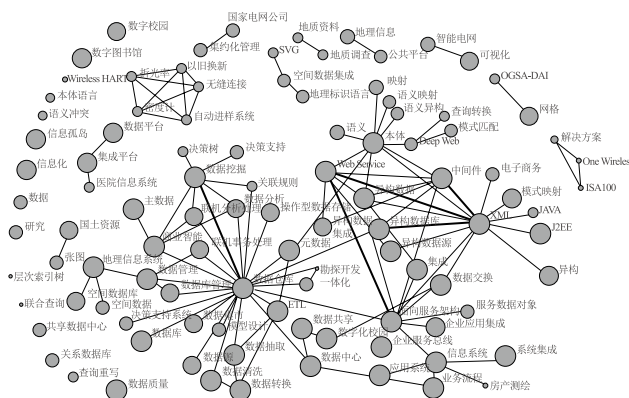


图9 2010-2014年间国内数据整合研究主题网络图

Service、数据仓库有很大的关系。从图中较弱的关系群或零星的数据点可以看出有关地理信息和空间信息(国土资源、地理信息、地质资料、房产测绘、地理标识语言和空间数据集成等)的研究明显增多,另外校园数据整合(数字校园、数字图书馆、数字化校园)和医院信息整合(医院信息系统)也开始得到学者的关注。

4 数据整合研究进展对比分析

从本文第三部分的分析发现,国外有关数据整合的研究起源于两个关系密切的概念GIS(地理信息系统)和遥感(Remote Sensing),并且有关这两者的研究一直是数据整合领域的热点,在划分的四个阶段中持续得到广大学者的关注^[6-10]。在2000-2003年段,除了有关地理信息系统的研究,关注度比较高的是Data Warehouse(数据仓库)和Data Mining(数据挖掘)及其之间关系的研究^[11-13]。而到了2004-2009年间Data Mining成为最受关注的研究热点之一,与之联系最为紧密的是数据挖掘技术的一个具体方法Clustering(聚类)^[14,15];此阶段最新出现的研究热点是Ontology(本体)^[16,17],与本体联系最为紧密的是Metadata(元数据)、Semantic(语义)、XML(可扩展标记语言)。在2010-2014年间除了Data Mining、GIS、Database等研究热点以外,有关数据整合的研究主要侧重于Bioinformatics(生物信息学)、gene expression(基因表达)和Cloud Computing(云计算)等方面^[18-22]。由以上的阶段分析可以看出,国外有关数据整合的研究正处在一个由基础技术研究过渡到应用数据整合发现更多数据中潜在价值的过程,现有的应用研究更多的侧重于公共事业方面;另外从其发展变化来看,有关数据整合的研究和信息技术的发展有很大的关系。

国内有关数据整合的研究热点和国外有些不同,在1991-1999年间主要是数据仓库、决策支持系统、联机分析处理和计算机辅助制造等方面,并且迄今为止,数据仓库一直是该领域的研究热点^[24-26];而2000-2003年段的研究则主要侧重于数据存储(数据仓库、数据集市、存储、Sybase等)和数据挖掘方面^[24,27];在2004-2009年间最突出的研究热点是以数据仓库-数据挖掘-商务智能三者形成的三角概念群,以及围绕XML形成的研究群Web Services、异构数据和中间件^[28-30];在2010-2014年间除了在第三阶段比较热的研究点以外,比较新的研究热点是面向服务架构、本体和地理信

息整合^[31-33]。由以上四个阶段的研究变化来看,国内有关数据整合的研究发展相对缓慢,更多侧重于基础层面,有关数据整合的应用则侧重于商业方面。

就国内外数据整合研究的对比分析来看,虽然国内数据整合文献数量的增长速度超过国外,但国内研究依然落后于国外,像有关本体的研究在国外于2004-2009年间已达到高峰,而国内于2010-2014年间才开始进行;国外运用云计算技术进行生物信息整合、基因数据整合时,国内才开始关注本体、地理信息整合等研究,而地理信息整合一直是国外学者研究的热点;国内持续得到广大学者关注的是有关数据仓库的研究,而国外有关数据仓库的研究在2000-2003年间得到关注之后就不再是数据整合研究的热点。

5 结语

综上所述,国内外有关数据整合的研究更多的偏向于数据整合的基础和技术研究,如数据仓库、数据存储、本体、可扩展标记语言、语义网络、元数据、异构数据等方面,而面向应用的研究相对来说还比较少,不过国内外都有向数据整合应用研究过渡的趋势,如生物信息整合、基因数据整合、GIS和面向服务架构等方面的研究。此外,国内学者已经对数据整合有了很高的关注度,但较国外而言,一直处于追赶和学习的阶段,并且有关数据整合高质量的研究文献比例偏小。产出更多高质量的研究成果并把更多的信息技术应用到数据整合研究中是国内学者下一阶段的方向。

参考文献

- [1] 张晗, 王晓瑜, 崔雷. 共词分析法与文献被引次数结合研究专题领域的发展态势[J]. 情报理论与实践, 2007, 30(3): 378-380.
- [2] 钟伟金, 李佳. 共词分析法研究——供词分析的过程与方式[J]. 情报杂志, 2008(5): 70-72.
- [3] 蔡立新. 比较管理研究的演化分析方法——范畴、意义及应用路径[J]. 比较管理, 2010(1): 014.
- [4] 陈权宝, 聂锐. 基于GPCA的高技术产业技术创新能力演化分析[J]. 中国矿业大学学报, 2005, 34(1): 117-122.
- [5] Hill H.A. Review and integration of physiological data[J]. Journal of Speech Disorders, 1944, 9(4): 289-324.
- [6] Ma Z, Hart M M, Redmond R L. Mapping vegetation across large geographic areas: Integration of remote sensing and GIS to classify

- multisource data[J]. Photogrammetric Engineering and Remote Sensing, 2001, 67(3): 295-308.
- [7] Ma Z, Hart M M, Redmond R L. Mapping vegetation across large geographic areas: Integration of remote sensing and GIS to classify multisource data[J]. Photogrammetric Engineering and Remote Sensing, 2001, 67(3): 295-308.
- [8] Yong L I, Shi-tai B A O, Ping Z. Research on data integration of ECDIS and GIS [J]. Science of Surveying and Mapping, 2007(4):052.
- [9] Karnatak H C, Shukla R, Sharma V K. Spatial mashup technology and real time data integration in geo-web application using open source GIS—a case study for disaster management[J]. Geocarto International, 2012, 27(6): 499-514.
- [10] Salleh S A, Hamid J R A, Ariffin I M. Investigation of potential integration of spectroradiometer data with GIS technology: The Spectro-GIS tools[C]//IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2014, 18(1): 012053.
- [11] Mcknight W. The integration of data warehouse data with campaign management[J]. DM Review, 2001, 11(11): 26-28.
- [12] Calvanese D, De Giacomo G, Lenzerini M. Data integration in data warehousing[J]. International Journal of Cooperative Information Systems, 2001, 10(3): 237-271.
- [13] You J, Dillon T, Liu J. An integration of data mining and data warehousing for hierarchical multimedia information retrieval[C]// Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on. IEEE, 2001: 373-376.
- [14] YatiY, Jain R C. Integration of data mining and operation research[J]. Ultra Scientist of Physical Sciences, 2009, 21(1): 311-313.
- [15] Pintér G, Madeira H, Vieira M, et al. Integration of OLAP and data mining for analysis of results from dependability evaluation experiments[J]. International Journal of Knowledge Management Studies, 2008, 2(4): 480-498.
- [16] BerkoQ. Use of ontologies in the processes of semantic data integration[J]. Bulletin of the National University Lvivska Polytechnika, 2009(650): 18-27.
- [17] Wang C, Lu J, Zhang G. Web ontology data matching for integration: method and framework[J]. International Journal of Web Information Systems, 2009, 5(2): 220-238.
- [18] A Seoane J, Aguiar-Pulido V, R Munteanu C, et al. Biomedical data integration in computational drug design and bioinformatics[J]. Current computer-aided drug design, 2013, 9(1): 108-117.
- [19] Goble C, Stevens R. State of the nation in data integration for bioinformatics[J]. Journal of biomedical informatics, 2008, 41(5): 687-693.
- [20] Rezola A, Pey J, Tobalina L. Advances in network-based metabolic pathway analysis and gene expression data integration[J]. Briefings in bioinformatics, 2014, 16(2): 265-279.
- [21] Fagan A, Culhane A C, Higgins D G. A multivariate analysis approach to the integration of proteomic and gene expression data[J]. Proteomics, 2007, 7(13): 2162-2171.
- [22] De la Prieta F, Rodríguez S, Bajo J. Data integration in Cloud Computing environment[C]//Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support. Atlantis Press, 2013.
- [23] 王大力. 关于万维网(WWW)分布式GIS的技术实现[J]. 东北测绘, 1999, 22(1): 39-41.
- [24] 张蒲生, 苏运霖. 基于政务数据仓库的数据集成及关键技术[J]. 计算机工程, 2003, 29(9): 179-182.
- [25] 林子雨, 杨冬青, 宋国杰. 实时主动数据仓库中面向需求的实时数据集成方法研究[J]. 计算机研究与发展, 2007(44): 435-440.
- [26] 彭耘, 汪丽华. 基于数据仓库技术的数字化校园数据集成研究[J]. 科技创业月刊, 2014, 27(2): 180-182.
- [27] 陈鹏, 吕卫锋, 单征. 基于网络的入侵检测方法研究[J]. 计算机工程与应用, 2001, 19(1): 44-60.
- [28] 马鸣, 赵轶超. 实时商务智能的框架及其技术分析[J]. 计算机应用与软件, 2009, 26(10): 130-132.
- [29] 邓国军, 傅彦. 利用XML解决数据集成中若干难题的应用研究[J]. 成都信息工程学院学报, 2006, 21(4): 513-518.
- [30] 席景科, 闫大顺. Web数据挖掘中数据集成问题的研究[J]. 计算机工程与设计, 2006, 27(8): 1366-1368.
- [31] 薛云. 现代企业中商务智能与知识管理的整合[J]. 中国流通经济, 2014, 28(5): 95-100.
- [32] 孟宇龙, 印桂生, 王慧强. 异构安全数据集成的本体建模方法研究[J]. 计算机工程与应用, 2010, 46(19).
- [33] 张庆福, 万麟瑞. 基于 SOA 的异构数据集成软件架构研究[J]. 计算机技术与发展, 2011, 21(5): 17-21.
- [34] Doyle FJ. Integration of photogrammetric and electronic data processing equipment for highway planning[J]. Photographic Science and Engineering, 1960, 4(2): 122-122.
- [35] Liddell DW. Integration and automatic fault location techniques in large digital data systems[J]. Communications of the Acm, 1962, 5(3): 139-139.
- [36] Li Y, Hedlund M, Kjellberg T. System integration for kinematic

- data exchange[J]. International Journal of Computer Integrated Manufacturing, 2015, 28(1): 87-97.
- [37] Ethier J, Curcin V, Barton A. Clinical Data Integration Model Core Interoperability Ontology for Research Using Primary Care Data[J]. Methods of Information in Medicine, 2015, 54(1): 16-23.
- [38] 白如江, 冷伏海. “大数据”时代科学数据整合研究[J]. 情报理论与实践, 2014, 37(1): 94-99.
- [39] 白晓亭. 大数据时代图书馆数据整合系统构建研究[J]. 图书馆建设, 2014, 240(6): 83.

作者简介

李伟华, 男, 1989年生, 硕士在读, 研究方向: 企业竞争情报、通讯作者, E-mail: 1039205534@qq.com。

郑彦宁, 男, 1965年生, 博士后合作导师, 研究馆员, 研究方向: 情报技术和方法、竞争情报。

刘志辉, 男, 1979年生, 博士, 副研究员, 研究方向: 情报分析方法、竞争情报、学科情报与战略情报研。

Analysis on the Research Progress of Data Integration at Home and Abroad

LI WeiHua, ZHENG YanNing, LIU ZhiHui

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: The rapid development of Internet technology enables enterprises and research work towards data-oriented, but the problem is scattered, isolated and difficult integration of different data. Data integration is also an important part of information science, so it's very important to find the difference between home and abroad about the research of data integration from the information science perspective. In this paper, co-word analysis was used to find the difference between home and abroad. The study found that the domestic research lags behind other countries, and the domestic research mainly focused on business intelligence and service oriented applications, but other countries paid more attention to geographic information systems, bioinformatics and gene expression.

Keywords: Data Integration; Thematic Analysis; Co-word Analysis;

(收稿日期: 2015-04-15; 编辑: 雷雪)