

大数据环境下知识发现研究的变化 及其发展趋向*

靳嘉林^{1,2} 王曰芬^{1,2}

(1. 南京理工大学经济管理学院, 南京 210094; 2. 江苏省社会公共安全科技协同创新中心, 南京 210094)

摘要: 随着信息技术的进步与知识经济的快速发展, 知识发现作为知识创新和知识服务的关键过程, 受到学术界的关注。知识发现以数据为起点, 因此大数据的兴起又为知识发现提出新课题。本文以CNKI和WOS为数据源检索文献, 以大数据被关注变化突增的2012年作为时间节点, 分别从国内、国外对比分析2012年前后的研究差异, 以探究大数据对知识发现研究的影响。进而, 通过对2012年以后知识发现文献的内容解析, 评述大数据影响下知识发现研究内容的变化与发展。结果发现, 大数据的出现使得国内知识发现研究方法研究的重点从数据挖掘技术转为数据准备技术和应用, 而国外的研究重点涉及知识发现的各个阶段, 国内需要加强知识发现的结果评估与解释阶段研究。

关键词: 知识发现; 大数据; 研究进展; 研究趋向; 对比分析; 内容分析

中图分类号: G350

DOI: 10.3772/j.issn.1673-2286.2018.05.010

目前知识发现的定义较熟知的是由Fayyad等^[1]在1996年提出, 指从数据集中提取有效的、新颖的、潜在有用的、可理解的模式的非平凡过程。模式是对数据集特征的描述, 识别出模式意味着为数据建立一个模型, 发现数据的内在结构, 产生数据集的高级描述。有效的指所发现的模式对于新的数据有一定的可信度; 新颖的指所发现的模式是以前未被挖掘出的新知识; 潜在有用的指所发现的模式能够为用户提供直接的经济价值或指导用户进行有意义的活动; 可理解的指所提取的模式容易被用户识别和掌握; 非平凡指知识发现过程是自动的或半自动的, 并有一定程度的智能性, 而仅发现所有数据的总和不能算作一个知识发现过程。

随着21世纪知识经济时代的到来, 人们对知识的需求变得更加迫切, 知识发现的研究也随之热门起来。数据是知识发现的基础与前提, 数据处理的理论、方法与技术的进步, 势必影响知识发现的过程与结果。1996年, 在学术研究中开始出现大数据术语, 2012年相

关研究逐渐呈现快速上升的趋势^[2]。大数据研究与应用的主要目的之一是知识发现, 为知识创新与知识服务提供支撑。因此, 大数据背景下, 知识发现的研究呈现怎样的变化趋势, 其发展态势如何, 是值得学界研究的一个重要课题。

1 大数据环境下国内外知识发现研究的概况及其对比

通过已有研究可知, 大数据概念是于2012年得到广泛认可并快速出现在各大学术交流会议与期刊中, 成为各学科的研究热点, 渗透到各行各业。因此, 为更清楚地了解大数据对于知识发现的影响, 本文将2012年作为研究变化的时间节点, 对比分析国内外知识发现研究的发文概况, 数据来源分别选取中国知网(CNKI)和Web of Science核心合集(WOS)。

从发文量看, 1992年国内首次出现明确的知识发

*本研究得到国家社会科学基金重大项目“面向知识创新服务的数据科学理论与方法研究”(编号: 16DZA224)和江苏省研究生科研创新计划项目“基于数据科学理论的科技情报知识库构建研究”(编号: KYCX17_0302)资助。

现研究,但在研究初期鲜有学者关注。国外知识发现研究略晚于国内,于1996年发表第一篇文章,但研究初期的关注度明显高于国内。进入21世纪,人们对知识的需求增加,如何从数据中得到知识并从知识中凝练新知识成为研究重点,这使得知识发现的发文量显著增加,国内外的发文量均在2005年达到峰值。随着知识经济热度的逐渐降低,知识发现的研究在经历短暂的高潮后缓慢回落。大数据环境带来新的技术和应用方向,使知识发现的研究趋势回暖,重新出现升温的态势,但新兴的内容仍在探索阶段,还未爆发足够的活力吸引大量学者加入,故国内外的发文均趋于平稳。从整体上看,国内外知识发现的研究热度变化几乎相同,但国外每年发文量均高于国内,国外研究知识发现的学者群体较国内多。

从研究方向看,大数据环境对国内外知识发现研究有不同的影响。在国内,受大数据环境的影响,图书情报与数字图书馆学科的研究比例明显升高,同时医学领域相关学科兴起,很多工学学科不再热门。而国外的知识发现研究有所不同,在大数据环境下,医学领域的研究逐渐消退,而其他领域对知识发现技术有更高要求的研究方向成为研究热点。造成这一现象的原因,一方面是国内外所处的社会环境不同,导致研究需求不同,国外对技术研究的需求高于国内;另一方面是国外该研究方向的学者数量较多,研究方向多样化发展得更为迅速。

从研究热点看,大数据环境对国内外知识发现研究的一个主要影响是使研究热点更加集中。2012年以前,知识发现研究与数据挖掘研究有大量重合部分,通常被归为一类。但从知识发现的实质来看,它是一个综合的知识活动和知识生产过程,涉及规律、策略和技术的集成,必须结合应用研究才能体现更大价值。因此,国内外知识发现的研究呈现一种从理论研究向应用研究转移的态势,但转移的程度和方向有所不同。对于国内知识发现研究而言,大数据受到广泛关注后,决策树、遗传算法、机器学习、文本挖掘等具体的理论方法不再是热点;而在数据驱动下,研究重点由如何应用知识发现的技术方法解决各行各业遇到的问题,逐渐向基于发现环境、面向领域和服务的综合应用研究转变,大数据、系统、可视化、图书馆、伤寒论等应用类关键词频现并成为研究热点;对于国外知识发现研究而言,除保持知识发现理论与技术研究的重点关注地位外,入侵检测、异常检测等应用类研究逐渐成为热点。

在大数据环境下,国内外的知识发现研究受到不同的影响。总体来说,2012年以前,知识发现的应用范围集中在计算机、工业、商业领域,着重研究相关方法技术和理论基础,然而数据的复杂性使得知识发现在处理更多领域的海量数据时难以发挥其作用,实践应用范围较窄;2012年后,大数据分析技术的应用帮助知识发现解决其数据准备阶段的局限性,在数据驱动下知识发现被用于解决更多领域复杂数据的问题,应用范围扩展到医药、科研、入侵检测等更多研究应用领域。

2 知识发现研究的主要观点与趋向分析

通过对CNKI和WOS中检索到的文献进行内容解析、阅读与归类,发现知识发现的研究主要集中在知识发现的理论、方法、应用三个方向。伴随大数据的快速发展,数据越来越丰富,而知识越来越贫乏,知识发现在三个方向上的研究也出现相应的改变。下文将针对知识发现的理论、方法、应用的主要观点与研究趋向进行评述和分析。

2.1 知识发现的理论

知识发现是一个集图书情报、计算机、数学等多个学科理论的研究领域,在大数据的影响下,原有理论难以支撑知识发现的技术及应用,因此,许多学者对大数据环境下知识发现的理论提出新见解,并对原有理论进行扩充。

段伟文^[3]将多元主义、视角主义、诠释学和能动者实在论应用到大数据知识发现理论中,认为大数据知识发现的实践是对数据的诠释,是复杂系统的自适应与自反馈过程;刘江玲^[4]认为知识发现系统应通过知识挖掘、数据分析等技术手段将资源整合、知识发现、信息推送融为一体,从而实现知识价值再造;李楠^[5]构建了以关联数据为底层支撑和逻辑控制,以知识发现为流程和结构的控制,以关联数据的应用功能为关键操作控制的基于关联数据的知识发现模型,延伸了知识发现研究的理论体系;Touzi等^[6]提出群集知识发现理论,可以支持更多的数据类型和分类算法;Espinoza-Molina等^[7]通过数据挖掘和知识发现方法推断地球观测传感器图像中隐藏的知识,提出复杂图像知识发现的理论;Zhao等^[8]从知识工程的观点出发,讨论人群踩踏事件的演化机制,并进一步提出基于粗糙

集理论的新型知识发现模型,以防止火灾抢险时发生人群踩踏事件。

总体上看,在大数据环境下,知识发现理论的研究重点在于扩充其理论体系,以计算机和数学为基础融合多学科的理论知识,结合关联数据、粗糙集,甚至神经网络、计算机视觉、自然语言处理等人工智能理论的应用,使知识发现理论能够支撑复杂的数据类型、不断优化处理技术及各领域的应用需求。

2.2 知识发现的方法

知识发现的基本过程有数据准备阶段、数据挖掘阶段、结果评估与解释阶段。知识发现是一种面向用户的服务,其中,数据准备阶段从用户需求入手;数据挖掘阶段结合用户需求与现有数据,使用数学或计算机方法进行知识发现;结果评估与解释阶段将知识发现结果展示给用户。大数据的兴起,打破了原有数据来源和结果展示的局限,扩充了知识发现的应用范围。在大数据环境下,知识发现的研究在优化数据挖掘阶段技术方法的同时,更多地从另外两个阶段探究知识发现的应用。

2.2.1 数据准备阶段

数据准备阶段对数据类型选择是知识发现在2012年前难以广泛应用的重要原因。随着大数据处理技术的发展,复杂的数据类型也可以使用合适的方法进行处理。在这一阶段,知识发现需要解决如何保证数据集模式可被识别的问题,在大数据影响下,这一问题应扩充成从获取数据集到模式可被识别的全过程。同时,伴随数据密集型研究范式的影响,数据驱动的知识发现研究不再是为达成一个目的而去寻找合适的、能处理的实验数据,而是对已有数据或获取的数据进行处理,从而达到知识发现的目的。因此,数据准备阶段显得尤为重要。

很多学者根据自身领域数据的特征,结合大数据技术和知识发现技术,对数据准备阶段的技术方法进行研究。如刘树林等^[9]根据中医经方的数据特征,将应用数学中的形式概念分析理论和属性偏序理论引入知识发现,使得大量的、不完全的、繁杂的中医数据能够有效地表达;吴剑锋^[10]优化大数据时代下面向知识发现的网络信息提取方法,以更好地利用、提取、转换、

过滤网络上庞大的结构化、半结构化和非结构化的复杂数据;Talía^[11]为解决知识发现过程中庞大数据的存储与分析问题,使用云计算对目前的数据挖掘云框架和分布式数据分析工具进行扩展开发;Castelli等^[12]提出利用几何语义遗传编程解决大数据知识发现的许多复杂数据问题;Galar等^[13]提出基于跨行业数据挖掘标准流程的故障诊断知识发现过程,解决异构数据源、可扩展数据结构方面的问题;Chen等^[14]开发了一种统一功能关联测量方法(BUFAM)来解决生物医学的“全功能频谱”知识发现中存在的异构知识源问题。

总体上看,大数据在带来海量数据的同时,也带来处理不同数据类型的思路。在大数据环境下,数据准备阶段逐渐得到学者们的关注,目前面对复杂的数据类型,知识获取与知识表示技术不再受到局限。适用于复杂网络结构的爬虫技术、面向多源异构可扩展数据的表示技术、处理庞大数据量的关联数据技术与分布式存储架构成为目前研究的重点,同时学者们根据自身研究数据特征和需求不断进行扩展与改进。

2.2.2 数据挖掘阶段

数据挖掘是知识发现的关键步骤,主要解决如何使用数学统计或计算机自动化方法识别数据集模式,并保证该模式是有效的、新颖的、潜在有用的、可理解的。根据用户需求,在数据挖掘阶段可采用的技术方法主要包括聚类法、分类法、决策树、粗糙集、关联规则、遗传算法、神经网络、机器学习、关联数据等。

2012年以前,国内数据挖掘技术研究较多,而2012年以后国内对于数据挖掘的研究集中在应用方面。如王莉^[15]提出一种基于粗糙集和模糊集的集成分类方法,优化了知识发现中数据挖掘阶段的分类方法。

在国外知识发现研究中,数据挖掘的技术研究始终是热点。如Cabrera等^[16]面对多变量时间序列数据提出知识发现的时间系列数据挖掘方法;Lomotey等^[17]认为,现有的数据挖掘技术是面向结构的存储设计,不适用于非结构化的数据样式,基于标签、过滤、关系图 and 自适应字典优化了数据挖掘技术;Ma等^[18]深入分析数据挖掘技术的瓶颈,并提出知识发现领域基于深度学习的挖掘模型体系架构;Jiang等^[19]对日志文件进行数据挖掘处理,为达到用户知识发现的目的,提出基于时间的聚类方法;Bai等^[20]对自组织特征进行优化得到批学习自组织特征,并用于知识发现的聚类分析。

知识发现过程中,数据挖掘技术实现的好坏通常直接决定知识发现结果的正确性,技术的优化才能促进知识发现产品的创新,大数据的出现对于优化数据挖掘技术的研究起到一定促进作用。如何将多学科的理论与方法融合,并针对多源异构数据的应用来深化研究具体的数据挖掘方法,特别是采用自然语言处理等优化数据挖掘技术,仍将是知识发现领域的研究重点。

2.2.3 结果评估与解释阶段

结果评估与解释阶段主要用于衡量知识发现过程,需解决数据集模式如何以合适的方式展示给用户。为解决这一问题,对结果的评估、可视化和推荐方法研究必不可少。如Braun等^[21]将社会网络分析用于知识发现,通过社交网络作图,为知识发现的可视化提供新途径;Sung等^[22]认为在数据经济环境下,知识发现的数据分析结果并非完全符合需求,因此,提出一种数据集质量评估措施来评估知识发现的结果;Bai等^[20]对自组织特征进行优化得到批学习自组织特征并用到知识发现的可视化中;Alvarado-Perez等^[23]综述了近年来可用于大数据的知识发现可视化技术、可视化系统和可视化模型,并根据数据挖掘的集成过程提出可视化技术的优化方法;Jiang等^[19]对日志文件进行数据挖掘处理,提出基于时间的推荐方法并用于用户知识发现。

结果评估是对知识发现过程好坏的判断,可视化和推荐是直接呈现给用户的知识发现结果展示方式,给用户更直观的感受。大数据为知识发现研究带来机遇,大力促进知识发现的技术与应用研究,但知识发现是一种面向用户的服务,研究不能脱离用户感受,结果的评估与展示十分重要。

从已有研究中得出,在大数据环境下,该阶段国外研究的重点已经集中在对现有的评估、可视化和推荐技术的优化上,而非已有方法的直接使用,优化的目标并非集中在某一技术方法上,说明这一阶段的技术方法均有广阔的优化空间,能够取得不错的研究进展。而在国内,并未发现明确提出对知识发现结果评估与解释阶段的研究,涉及该阶段的研究多是对已有技术方法的直接应用,对此阶段的关注还有待提高。

2.3 知识发现的应用

理论技术方法研究最终都将归于应用,知识发现

也不例外。知识发现是一个智能化、自动化的过程,其目标是帮助人们解决数字信息时代的信息过载问题^[24]。特别是大数据处理技术的兴起使得知识发现能够更好地应用于各个领域,解决领域内的问题。对此,国内外学者展开了相关的研究。

陈聪等^[25]将大数据挖掘和知识发现技术应用于辅助智慧城市的综合发展;赵瑞雪等^[26]设计了一种大数据环境下的农业知识发现服务体系,从农业综合科技数字知识存储构建、基于深度融合知识挖掘分析等方面阐述农业知识发现服务的研究;殷雯雯等^[27]将大数据技术应用于城市配电网中,提出基于城市配电网易购数据的知识发现和系统优化概念;Holzinger等^[28]将知识发现方法应用于生物医学研究领域,认为知识发现和数据挖掘的方法非常适合生物医学研究者;Renu等^[29]采用知识发现和数据挖掘算法,通过怀卡托环境知识分析(WEKA)接口进行时间研究分析,为装配生产线规划提供决策支持;Huang等^[30]通过大数据知识发现的LDA算法检测网络中异常入侵现象;Mishra等^[31]提出一种面向认知的物联网大数据框架,用于大型工业自动化应用的知识发现系统;Zhang等^[32]基于数据库和大数据知识发现分析网络舆情与知识发现的影响,将知识发现用于智慧城市建设中网络舆情的管理。

总体上看,在大数据的支撑下,知识发现的应用范围逐渐扩展。从数据来源上看,知识发现的应用范围从仅支持关系数据库管理系统、联机分析多维数据集等结构数据集变为可以支持多维、异构、无序、碎片式等复杂数据集,如网络日志、电子邮件、社交媒体等;从研究领域上看,知识发现的应用范围在计算机、工业、商业的基础上增加其他领域,如医学、农业等;从用户服务类型上看,知识发现的应用范围从为企业等集团型用户提供标准化服务扩展为包括个体用户在内的多种用户的个性化服务。

3 总结

本文通过对国内外2012年前后研究概况的对比分析可以发现,大数据对国内知识发现发文量的影响不大,但是对发文内容影响较大,最明显的是图书情报领域对知识发现的高度关注和医学领域知识发现的快速兴起。通过对大数据环境下国内外知识发现文献的内容分析可以发现,国内外都有部分文献在研究知识发现的技术方法,并将其应用于各个领域。同时,国外对

于知识发现的技术方法研究较全面,而国内知识发现的技术方法研究有不同的侧重,2012年前注重数据挖掘方法,2012年后注重数据准备相关方法,但是对于结果评估与解释相关方法的研究存在局限。知识发现作为面向用户的知识服务,其结果评估与解释阶段的研究将是必不可少的。

目前,知识发现领域的研究正处于积蓄力量的阶段,扩展知识发现各阶段的研究将为知识发现研究的发展带来充足活力。理论研究应着重与其他学科理论的融合,而非多学科理论的直接调用,为处理大数据环境下的复杂数据提供理论支撑。方法研究的重点应从三个阶段分别入手:数据准备阶段可加强爬虫技术、多源异构可扩展数据表示技术、关联数据分布式存储技术等研究;数据挖掘阶段应改变现有技术直接利用的问题,针对不同数据优化数据挖掘算法,深入研究计算机视觉、自然语言处理等人工智能技术,探究将这些技术移植到知识发现研究的可行性,并根据知识发现研究的特性对其进行优化;结果评估与解释阶段应当予以重视,大量的评估、可视化、推荐等方法在知识发现体系下可以得到深度优化,这将是一个值得研究的领域。应用研究受大数据的影响,研究范围得到扩展,个性化的、诸多领域的、复杂数据的应用都将成为可能。

由于知识发现存在很多细分领域,难以全部涵盖,本文数据的选取可能未覆盖所有知识发现的相关文献。但是,本文从知识发现的理论、方法和应用角度,结合大数据对相关研究的影响,力争对研究进展、主要观点与发展趋向进行述评和分析,为掌握知识发现研究的现状与发展提供参考与借鉴。

参考文献

- [1] FAYYAD U, PIATETSKY-SHAPIO G, SMYTH P. The KDD process for extracting useful knowledge from volumes of data [J]. Communications of the ACM, 1996, 39 (11): 27-34.
- [2] 王曰芬,谢清楠,宋小康.国外数据科学研究的回顾与展望[J].图书情报工作,2016,60(14):5-14.
- [3] 段伟文.大数据知识发现的本体论追问[J].哲学研究,2015(11):114-119.
- [4] 刘江玲.面向大数据的知识发现系统研究[J].情报科学,2014(3):90-92.
- [5] 李楠.基于关联数据的知识发现研究[D].北京:中国农业科学院,2012.
- [6] TOUZI A G, ALOUI A, MAHOUCI R. Cluster_KDD: a visual clustering and knowledge discovery platform based on concept lattice [C] // International Conference on Advances in Swarm Intelligence. Springer Berlin Heidelberg, 2012: 127-136.
- [7] ESPINOZA-MOLINA D, DATCU M. Data mining and knowledge discovery tools for exploiting big earth-observation data [J]. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, XL-7/W3 (7): 627-633.
- [8] ZHAO R Y, LIC L, LING W Q, et al. A rough set-based knowledge discovery model for crowd evacuation in a fire emergency [C] // International Conference on Advanced Management Science and Information Engineering (AMSIE 2015). Destech, 2015: 37-42.
- [9] 刘树林,徐笋晶,李日东,等.基于属性偏序原理的仲景经方咳嗽治疗规律知识发现[J].燕山大学学报,2014(5):455-459.
- [10] 吴剑峰.大数据时代面向知识发现的网络信息提取方法研究[D].合肥:安徽理工大学,2016.
- [11] TALIA D. Making knowledge discovery services scalable on clouds for big data mining [C] // IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services. IEEE, 2015: 1-4.
- [12] CASTELLI M, VANNESCHI L, MANZONI L, et al. Semantic genetic programming for fast and accurate data knowledge discovery [J]. Swarm & Evolutionary Computation, 2015, 26: 1-7.
- [13] GALAR D, KANS M, SCHMIDT B. Big data in asset management: knowledge discovery in asset data by the means of data mining [C] // Proceedings of the 10th World Congress on Engineering Asset Management (WCEAM 2015). Springer International Publishing, 2016: 161-171.
- [14] CHEN H, WEI C, LIU C, et al. Relational network for knowledge discovery through heterogeneous biomedical and clinical features [J]. Scientific Reports, 2016, 6 (1): 29915-29927.
- [15] 王莉.基于粗糙-模糊集成的分类知识发现[D].南京:南京大学,2013.
- [16] CABRERA F, SHIN R, CONCHA D, et al. Temporal knowledge discovery in big BAS data for building energy management [J]. Energy & Buildings, 2015, 109 (4): 75-89.
- [17] LOMOTY R K, DETERS R. Towards knowledge discovery in big data [C] // IEEE, International Symposium on Service Oriented System Engineering. IEEE Computer Society, 2014:

- 181-191.
- [18] MA Y, TAN Y, ZHANG C, et al. A data mining model of knowledge discovery based on the deep learning [C] // Industrial Electronics and Applications. IEEE, 2015: 1212-1216.
- [19] JIANG Y, LI Y, YANG C, et al. Reconstructing sessions from data discovery and access logs to build a semantic knowledge base for improving data discovery [J]. ISPRS International Journal of Geo-Information, 2016, 5 (5): 54.
- [20] BAI Y, IWASAKI Y, KANAYA S, et al. A novel bioinformatics method for efficient knowledge discovery by BLSOM from big genomic sequence data [J]. Biomed Research International, 2014, 2014 (5): 1-11.
- [21] BRAUN P, CUZZOCREA A, LEUNG C K, et al. Knowledge discovery from social graph data [J]. Procedia Computer Science, 2016, 96: 682-691.
- [22] SUNG A H, RIBEIRO B, LIU Q. Sampling and evaluating the big data for knowledge discovery [C] // International Conference on Internet of Things and Big Data. SCITEPRESS, 2016: 378-382.
- [23] ALVARADO-PEREZ J C, BOLANOS-RAMIREZ H, PELUFFO-ORDONEZ D H, et al. Knowledge discovery in databases from a perspective of intelligent information visualization [C] // Signal Processing, Images and Computer Vision. IEEE, 2015: 1-7.
- [24] 王敏, 张志强. 图书情报领域知识发现研究文献内容分析 [J]. 现代图书情报技术, 2008 (2): 66-68.
- [25] 陈聪, 张国惠, 马晓磊, 等. 利用大数据挖掘和知识发现技术辅助智慧城市发展 [J]. 大数据, 2016, 2 (3): 39-48.
- [26] 赵瑞雪, 鲜国建, 寇远涛, 等. 大数据环境下的农业知识发现服务探索 [J]. 数字图书馆论坛, 2016 (9): 28-33.
- [27] 殷雯雯, 彭晨, 杜大军. 大数据时代城市配电网异构数据的知识发现与优化研究 [J]. 工业控制计算机, 2015 (11): 134-135.
- [28] HOLZINGER A, ZUPAN M. KNODWAT: a scientific framework application for testing knowledge discovery methods for the biomedical domain [J]. BMC Bioinformatics, 2013, 14 (1): 191.
- [29] RENU R S, MOCKO G, KONERU A. Use of big data and knowledge discovery to create data backbones for decision support systems [J]. Procedia Computer Science, 2013, 20: 446-453.
- [30] HUANG J, KALBARCZYK Z, NICOL D M. Knowledge discovery from big data for intrusion detection using LDA [C] // IEEE International Congress on Big Data. IEEE, 2014: 760-761.
- [31] MISHRA N, LINC C, CHANG H T. A cognitive adopted framework for IoT big-data management and knowledge discovery prospective [J]. International Journal of Distributed Sensor Networks, 2015: 6.
- [32] ZHANG C, HE L, MAO Y, et al. Knowledge discovery of network public opinion in the concept of smart city [C] // Industrial Electronics and Applications (ICIEA), 2015 IEEE 10th Conference on. IEEE, 2015: 1197-1202.

作者简介

靳嘉林, 男, 1992年生, 博士研究生, 研究方向: 知识发现与数据挖掘, E-mail: jinjialin9219@163.com。

王曰芬, 女, 1963年生, 教授, 博士生导师, 通信作者, 研究方向: 知识挖掘与知识服务、知识工程与知识管理, E-mail: yuefen163@163.com。

The Change of Knowledge Discovery Research and Its Development Trend in Big Data Environmental

JIN JiaLin^{1,2} WANG YueFen^{1,2}

(1. School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China;

2. Jiangsu Social Public Security Technology Cooperative Innovation Center, Nanjing 210094, China)

Abstract: With the progress of information technology and the rapid development of knowledge economy, knowledge discovery, as the key process of knowledge innovation and knowledge service, is concerned by academic circles. Data is the starting point of knowledge discovery, so the rise of big data has put forward a new topic for knowledge discovery. In this paper, CNKI and WOS retrieval are used as data sources, and 2012, which is highlighted by big data, is used as a comparative time node. The paper compare the differences before and after 2012, respectively from two aspects at home and abroad, in order to explore the impact of big data on knowledge discovery. Then, through reading and classifying, the paper do content analysis for the knowledge discovery literature after 2012, and review the change and development of the specific research contents of knowledge discovery under the influence of big data. Through the comparative analysis and content analysis, this paper found that the emergence of big data made the key of research on domestic knowledge discovery method from data mining to data preparation and application, while foreign research focus on all stages of knowledge discovery. Domestic research need to strengthen the results evaluation and interpretation stage research of knowledge discovery.

Keywords: Knowledge Discovery; Big Data; Research Progress; Research Trend; Comparative Analysis; Content Analysis

(收稿日期: 2017-12-25)