

# 大数据背景下的档案资源数据挖掘研究

赵志强

中铁二院工程集团有限责任公司档案馆

**摘要:**随着大数据时代的到来,档案资源的数字化和信息化程度不断提高,尤其是档案数据挖掘技术的应用,让企业的档案管理更加规范,能有效地将用户和档案数据连接起来,从而实现双向控制,极大地促进了企业档案数据的管理效率。然而如何挖掘这些数据并应用于档案资源数据管理中,成了档案管理者们的关注焦点。本文综述了大数据背景下的档案资源数据挖掘研究的最新进展和发展趋势。首先介绍了档案资源数据挖掘技术的基本特征和类型;其次阐述了档案资源数据挖掘的现状;在此基础上,针对档案资源数据挖掘的特点和难点,讨论了管理方式、管理者的管理意识等问题,最后提出了大数据背景下的档案数据挖掘技术的相关策略,希望能为其他档案管理者提供参考和启示。

**关键词:**大数据背景;档案数据挖掘;相关技术;现状;对应策略

**【DOI】** 10.12252/j.issn.2096-6288.2023.01.232

## 引言

近年来,随着各行各业信息化的推进和大数据技术的不断发展,档案资源在数字化转型方面也得到了更多人的重视。档案资源是一个国家、地区或者企业的历史、文化、经济等方面的重要证据,包含了许多有价值的信息和知识。目前档案资源的数字化建设一直以来都存在着诸多的问题和挑战。首先,许多档案资源还处于纸质状态,没有经过数字化加工处理;其次,由于档案资源来源广泛、内容复杂,这些数据需要经过整理、编目、标引等工作才能进行存储和使用;再次,档案资源中蕴含着大量的信息和知识,但如何从这些数据中挖掘出有价值的信息和知识还需要进一步研究和探索。此外,档案资源数据挖掘的应用领域也非常广泛。例如,在历史研究、社会学研究、文化研究、政治研究等领域,档案资源数据挖掘已经被广泛应用。通过对档案资源中的数据进行深入挖掘和分析,可以得到更加准确、全面的研究结果,从而促进相关领域的发展和进步。在当前档案数字化建设的大趋势下,各个企业对档案资源进行数字化转型已经成为必然趋势,而如何更好地利用这些数据,挖掘出其中的有价值信息和知识,也成了当前研究的重要方向。

## 一、大数据技术的基本特征

### (一) 信息庞大

这指的是大数据技术所涉及的数据量十分巨大,远远超过传统数据库技术能够处理的规模。如今,各种业务应用和社会互动行为都会产生大量的数据,例如物联网、金融交易等。这些数据量巨大的信息源不断涌现,使得数据总量呈爆炸式增长。数据信息总量庞大带来的是海量数据的处理和存储难题,需要采用分布式计算和

存储技术、云计算等高级技术进行数据管理。

### (二) 信息多样化

是指大数据技术所涉及数据的多样性,数据来源各种各样,包含结构化、半结构化和非结构化数据。如物理信息系统信息、国家安全信息、国家铁路信息等。此外,大数据还利用特殊的技术手段对非结构化的日志文件、网络流量等一些无法用传统数据库存储的数据形式进行提取和处理。

### (三) 高效性

这是指大数据技术在处理海量数据时具有高度的效率和速度。传统的数据处理方式会受到硬件设备等因素的限制,处理大量数据需要花费很长时间。而大数据技术则采用分布式计算、并行计算等技术手段,能够同时处理多个任务,提高数据处理效率。此外,大数据技术还采用了实时处理技术,能够快速响应和处理大规模的数据流量,能够满足实时性强的业务需求。

## 二、档案资源数据挖掘技术的基本类型

### (一) 分类技术

分类技术是指通过对已有数据进行分析 and 处理,将数据划分到不同的类别中,以便更好地进行管理和利用。分类技术可以帮助企业更好地理解数据,从而为其决策提供更好的支持。其主要作用在于对数据进行整合和概括,将相似的数据划分到同一类别中。分类通常采用分类树算法,通过建立一个树形结构来描述数据特征,然后根据数据的属性进行分组。在档案资源中,分类技术可以用于对档案文献、档案资料、档案案卷等信息进行分类。例如,可以将文献按照出版时间、作者、主题等方面进行分类;将资料按照制作时间、载体种类、主题等方面进行分类;将案卷按照档号、机构、档

案类型等方面进行分类。通过分类技术,企业可以更好地管理其档案资源,便于用户进行信息检索和利用。

## (二) 聚类技术

聚类技术是档案资源数据挖掘技术中的一种基本方法,主要用于对档案数据进行分类和聚集。它通过寻找数据集中相似的数据,并将其自动分组,从而为数据管理和利用提供更好的支持。聚类技术的核心在于寻找相似性。它会根据数据之间的距离、密度、相似度等因素来确定相似性的程度,然后将相似性较高的数据划分到同一个组别中。常见的聚类算法包括K-means、DBSCAN、层次聚类等。其中,K-means算法是最常用的一种算法,它基于样本之间的距离来计算相似性程度,并且通过质心迭代的方式,实现对数据的分组。在档案资源中,聚类技术可以用于对档案文献、档案资料、档案案卷等信息进行分类。例如,可以将文献按照出版时间、作者、主题等方面进行聚类;将资料按照制作时间、载体种类、主题等方面进行聚类;将案卷按照档号、机构、档案类型等方面进行聚类。这样,聚类技术可以帮助机构更好地管理和利用其档案资源。因此企业可以更容易地发现数据的特征和规律,发现隐含的信息,并提供更加精确、全面的查询结果。

## (三) 关联技术

关联技术是指通过对不同档案资源中的各种数据进行分析和挖掘,发现它们之间的相关性或关联性,并从中提取出有价值的信息。这是一种非常有用的技术,可以为机构提供更加全面、准确的档案资源管理和利用方案。其主要包括两种算法:关联规则算法和序列模式算法。其中,关联规则算法是最常用的算法之一。它可以发现不同数据之间的关联规律,体现出几件事情中的联系和规律。序列模式算法则可用于发现时间序列数据中的模式和趋势,例如可以分析企业过去的档案推广活动的周期性和趋势。企业可以通过挖掘不同档案间的关联,了解到某个时期的相关事件、人物和社会背景等信息。

## 三、档案资源数据挖掘技术的现状

### (一) 管理方式落后

首先是管理手段落后。数据挖掘技术需要大量的数据支持和分析工具,但是目前部分企业的档案管理系统仍然处于一个相对落后的状况,难以提供足够的技术支持,也难以实现对大数据的自动化分析和挖掘。此外,有些企业虽然部署了数字化档案管理系统,但是由于系统安装和维护成本高昂,以及工作人员的技术素质不高等问题,导致系统无法充分发挥作用。其次是管理流

程落后。数据挖掘技术需要从数据源开始,通过数据清洗、预处理、建模等环节,最终得到有意义的结果。而在传统的档案管理中,由于流程繁琐、耗时长等问题,往往导致管理效率低下和数据质量不佳。因此,在实际应用中,管理者需要重新审视管理流程,将数据挖掘技术纳入整个流程中,以实现信息的快速获取、准确分析和智能应用。

### (二) 档案管理者的管理意识落后

首先,管理者缺乏数字化和智能化建设的认识。在传统档案管理方式主要依靠人工和手工操作,管理过程存在很多瓶颈和缺陷,例如信息获取不及时、处理速度慢等。而数据挖掘技术可以大幅度提高信息处理效率和准确性,但是在实际应用中,许多企业仍然没有树立数字化、智能化管理理念,缺乏前瞻性的思维 and 创新能力。其次,管理者缺乏挖掘数据的意识。档案资源数据挖掘技术需要从数据源开始,通过数据录入、预处理等环节,最终得到有意义的结果。然而,在实际应用中,很多管理者并没有意识到档案资源中蕴含着大量有价值的信息,因此对于数据挖掘技术的应用缺乏理解和认可。

## 四、企业档案资源数据挖掘策略

### (一) 加强企业的内部管理

(1) 建立并完善企业档案管理制度:制定企业《电子文件归档范围及保管期限》《磁带档案管理与保护规定》等档案管理相关制度,建立完善的档案管理工作体系。

(2) 优化档案业务流程:根据企业完善内控体系的最新要求,结合企业实际,梳理科技档案业务流程,制定科技档案归档、利用、销毁管理流程及相关管理流程说明。

(3) 开发档案管理系统:按照《企业数字档案馆(室)建设指南》和《企业电子文件归档和电子档案管理指南》等文件要求,提出综合档案管理系统档案收集、档案整编、档案借阅、档案统计等各功能模块的需求,完成综合档案管理系统定制开发。

(3) 深化企业档案信息化建设:首先企业应该将纸质档案进行数字化处理,通过扫描、OCR文字识别、图像处理等技术手段将纸质档案转化为电子档案,并建立数字化档案管理系统,依托云计算、大数据、物联网等现代信息技术,实现企业档案信息的数字化、网络化、智能化,从而提高档案管理的效率和质量,实现档案的在线查询、检索、借阅等功能。此外不仅要负责档案馆网页的日常维护和更新,还要加强维护管理档案管

理系统。同时在企业档案信息化建设过程中,加强安全保密措施非常重要。通过加密、备份等技术手段,保证企业档案信息的安全性和机密性,并完成综合档案管理系统等保测评。

## (二) 利用大数据技术实现大数据挖掘的精细化

(1) 加强档案资源挖掘。首先要建立完善的档案资源管理系统。档案资源管理系统是进行资源挖掘的基础,需要建立详细的资源清单和档案目录,记录每一个档案文件的内容、来源、时间等信息。通过对档案资源的建立和管理,可以更加有效地发掘档案资源中的有价值的信息。其次是构建完善的档案资源数据共享平台。可以以省级档案资源库为基础,构建一个全面的、规范的、科学的数据共享平台,加强本省各个档案馆之间的联系,形成一个资源数据互通的关系网,及时接受和分享最新的档案数据,为用户提供更好的档案服务平台。同时引入先进的技术手段如图像识别、语义分析、机器学习等,可以有效提高挖掘效率和准确性。

(2) 加强档案数据挖掘。档案数据挖掘是指利用各种数据挖掘技术对档案数据进行分析 and 处理,以获取有价值的信息和知识。首先,要对档案数据进行全面、准确的收集、整理和存储,保证数据质量和可靠性。要注重数据的标准化、格式化、去重、清洗和修复等过程,避免数据孤岛,以保证数据的一致性和可比性。另外,还要利用数据质量工具对数据进行监控和评估。其次可以对用户的搜索记录和浏览记录进行收集和分析,利用大数据技术将搜索率较高的档案数据进行优化,将其以关键词的形式呈现在搜索栏下面,减少用户搜索时间,提高搜索的准确性和用户的满意度。在档案数据挖掘中,还可以引入机器学习算法,如决策树、支持向量机、神经网络等,对档案数据进行分类、聚类、回归等分析,从中发现规律并做出预测,挖掘出其中潜在的价值信息。同时可以将数据上传到云端,云计算平台具有强大的计算能力和存储能力,利用其进行分布式计算和多线程处理,以提高档案数据挖掘的效率和精度。

## (三) 通过综合性策略实现档案数据挖掘的深度化

(1) 加强以人为本的档案用户管理。首先利用大数据技术进行较为全面的数据搜集与整理工作,从大量的数据中挖掘出有用的信息,建立相应模型和算法,识别出客户的需求,进行有效的分析。同时对客户的特征进行分类,分析客户的搜索习惯、兴趣爱好等方面,为后续的关系管理提供数据支持。接着基于大数据技术,建立个性化服务体系,为档案用户提供定制化服务。根据用户的需求和偏好,提供不同的服务方案,例如,为

用户提供个性化的档案查询服务、为用户提供专业的咨询与答疑、为用户提供预约服务等,以满足用户不同层次、不同需求的服务要求。

(2) 加强档案数据采集层管理。第一,在建立数据采集层之前,需要明确采集数据的目标和流程。确定需要采集的数据类型、采集方式和频率。同时,制定适当的数据处理流程和记录方式,以确保数据的准确性和完整性。第二,设立质量控制机制。通过检查、审核和验证采集数据的过程来确保数据的质量,制定并实施合适的审核和验证程序,以确保数据的准确性和完整性。第三,规范化数据管理过程。采用标准化格式、规范化方法和结构化存储方法,以确保数据的一致性和可操作性。此外,建立数据分类和索引体系,以便快速检索和利用档案数据。第四使用数字化技术。采用数字化技术,如扫描、OCR等来构建一个智能化的档案管理系统。该系统将大大提高数据处理速度、节省人力成本,并增加数据可靠性和安全性。同时对企业所属各单位、子公司、项目部生产管理人员提供培训和指导,以确保他们了解采集流程和技术,并遵循国家、企业制定的相应的文件标准和规定,从而确保数据的准确性和一致性。

## 结语

在大数据时代,档案资源的数据挖掘研究日益受到重视。本文综述了当前档案资源数据挖掘研究的发展现状和趋势,以及所面临的问题和挑战。我们发现,随着数字化技术的普及和应用,档案资源的数据量不断增加,企业档案管理工作者还需要进一步深入研究档案资源数据挖掘的方法和技术,通过建立健全数据共享机制,加强企业的内部管理和档案数据的精细化与深度化,以提高档案资源的利用价值和服务水平,同时保障档案资源数据挖掘工作的顺利进行。

总之,大数据背景下的档案资源数据挖掘研究在中国具有广泛的应用前景和重要的研究意义。未来,企业可以通过整合各种技术手段、提升档案资源数据质量、建立标准规范等方式,提高档案资源的数字化转型水平和数据挖掘能力,为人们创造更加便利和高效的服务。

## 参考文献

[1] 左建安,陈雅.大数据时代的科学数据共享模式研究[J].2021(2014-3):32-35.

[2] 戴柏清,秦顺.国家文化数字化战略下档案馆发展向度与实践路径[J].北京档案,2022(8):5.

作者简介:赵志强,1975年6月17日,男,籍贯:江苏省徐州,学历:本科。