数据挖掘国内研究综述

杨小娟
（云南师范大学信息学院，昆明 650000）

摘　要：随着大数据时代的到来，数据挖掘也逐渐变成研究的热点问题，并应用于各个领域中。采用文献分析法和内容分析法，基于中国知网对数据挖掘相关文献检索，筛选出国内相关文献，并对其进行归纳整理，从概念简述、挖掘过程、主要方法、研究现状及发展趋势 4 个方面进行阐述，以期为今后开展相关研究提供参考借鉴。

关键词：大数据；数据挖掘；算法
DOI:10.16184/j.cnki.comprg.2020.08.041

基于文献研究方法，以“数据挖掘”为主题词，从中国知网文库中对相关文献进行检索，检索结果表明，相关文献共计 80862 篇，其中国内相关文献共计 78180 篇，随之利用中国知网计量可视化分析工具对所检索的全部文献进行可视化分析处理，结果表明相关文献主题关键词主要集中于数据挖掘技术、关联规则、数据仓库、大数据、决策树。因此采用内容分析方法，对文献进行分类梳理，并从概念简述、挖掘过程、主要方法、研究现状及发展趋势 4 个方面来对数据挖掘进行归纳阐述。

1　概述

数据挖掘兴起于 1989 年，又称数据库中知识发现。是多门学科知识融会贯通的产物，其中包括机器学习、数据库应用技术、统计学、人工智能等多个学科领域的研究成果[1]。

数据挖掘因此被定义为利用机器学习、统计学习等相关方面的知识和技术，从海量数据中整理、归纳、规律发现高价值模型或数据的手段，提取出新颖的、有效的、潜在有用的并且可被理解的模式处理过程[1]。如图 1 所示。

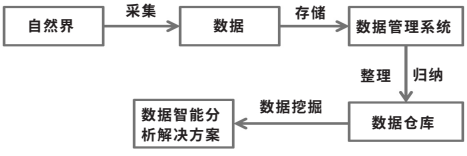

图 1　数据挖掘含义

2　挖掘过程

2.1　明确挖掘主题

在进行数据挖掘之前，先明确自己的数据导向，确定数据挖掘的方向及范围，再实施数据挖掘，以规避数据冗余、数据偏差等问题，避免盲挖。

2.2　数据处理

数据处理环节是数据挖掘过程中的重要环节，只有保证数据的准确有效性，才能保证数据挖掘的有意义性，其中此环节共分为 3 个小环节，分别是数据选择、数据预处理、数据转换。

（1）数据前期准备。根据研究主题收集相关数据，将收集后的数据进行归类整理，剔除与主题无关或偏差较大的数据，留下主题相符的数据。

（2）数据处理。将整理好的数据进行二次处理，对空白字段、无意义数据进行删除，保证所留下的数据都具有有效性。

（3）数据转换。将保留下来的有效数据根据所研究的主题目标进行聚类处理，以满足数据挖掘格式需求，是数据挖掘的先前条件。

2.3　数据挖掘

数据挖掘是对数据进行实质性挖掘，再根据主题选择适合该数据研究的算法，然后对数据实施挖掘工作，这一环节是数据挖掘工作的核心环节。

2.4　数据分析

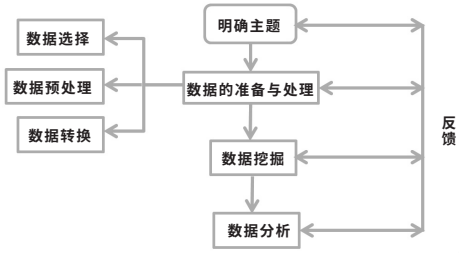

图 2　数据挖掘过程

作者简介：杨小娟（1995–），女，硕士，研究方向：数据分析。

数据挖掘工作结束后，最后一步根据所挖掘出来的数据结果对此研究进行阐述说明，其主要作用是确定知识的模式模型是否有效以便发现更加有意义的知识模型。如图 2 所示。

3 主要方法

3.1 决策树

决策树是数据挖掘的主流方法，以树形形式将数据决策与数据分类过程清晰描述，这种算法相对较简单、直观、易理解。在不同场景中生成的决策树也会不一样，所以决策树也会被称为分类树、回归树等。决策树数据挖掘方法经典算法主要为 ID3 算法和 C4.5 算法。

（1）Cart 算法。是简单的二叉树算法，它往往被运用于简单数据中，生成结构较简单的二叉树。

（2）ID3 算法。ID3 算法是决策树算法中相对较早的算法，它在数据信息基础上通过一系列规则找出树中每一个节点所代表的属性，以算法中的熵为分类依据，将数据最终生成决策树的形式。

（3）C4.5 算法。C4.5 算法在 ID3 算法的基础上进行优化改进，此算法使用信息增益或者熵来优化决策树节点划分的过程，修善决策树，使决策树更加友好。

3.2 聚类分析

聚类分析本质上是根据研究主题，找出数据的分类依据，并根据此依据对数据进行分类处理，将数据细化为不同类型的数据集合，并保证每个集合中的数据都具有相似性，不同集合之间又存在着差异性，再利用数据可视化技术将其表现出来，并友好地展现给用户，即称为聚类分析。其主要算法为 K-means 算法，这一算法的突出优势在于原理简单、应用高效，非常适合对规模较大的数据进行处理，在很多领域取得了较好的应用效果，包括：数据分析、个性化推荐、数据分类、图像识别等[2]。

3.3 关联规则分析

关联规则分析是数据挖掘工作中较为常用的方法之一。事物与事物之间存在着相互之间的关系，而这种关系称之为关联，关联规则指得是事物与事物之间潜藏的关系规则，而关联规则分析则指的是在事物与事物之间查找和分析那些所内设定值关联规则之间信息的过程。其主要算法为 Apriori 算法，它是一种最有影响的挖掘单维、单层、布尔关联规则频繁项集的算法。Apriori 算法虽然可以解决相应数据关联规则的分析，但是它还存在着一定的缺陷，随之 J. Han 等人提出了 FP-Growth 算法，以弥补 Apriori 算法产生候选项集的缺陷。

3.4 支持向量机

支持向量机（SVM）,是一种二分类模型，基本模型是定义在特征空间上间隔最大的线性分类器，间隔最大时它有别于感知器。核心理念是支持向量样本会对识别的问题起关键性作用，支持向量也就是离分类超平面最近的样本点，而这个分类超平面正是支持向量分类器，通过这个分类超平面实现对样本数据一分为二。如图 3 所示。

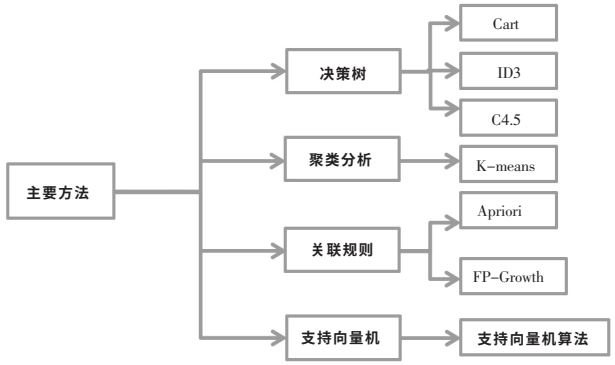

图 3

4 研究现状及发展趋势

4.1 研究现状

在研究过程中以课题的关键词“数据挖掘”，在知网中进行了相关文献检索，阅读了解，归纳发现国内数据挖掘兴起时间稍晚于国外，现阶段正处于发展期，相比国外数据挖掘技术还不成熟。其中国内数据挖掘研究人员主要集中于各大高校，其他企业相对较少，且各大高校此主题研究项目都基于政府支持下开展。

根据相关文献分析，其最新研究方向主要表现为：数据分析、数据风险评估及质量优化、预警模式构建、学习行为分析及教学评价、数据模型的构建与评价、粗糙集和模糊集理论的对比研究、智能专家系统的构造、中文文本挖掘、数据挖掘算法研究等方面。

4.2 发展趋势

4.2.1 多模态数据挖掘

从数据挖掘的对象来说，数据挖掘后期多会偏向多模态数据挖掘。因为就当前来看大部分的数据挖掘都是针对结构化数据进行挖掘的，但大数据时代背景下，非结构化数据占据主流，如果从这些非结构化数据中挖掘出隐藏信息，将是未来大数据领域研究和实践的重点。

4.2.2 数据挖掘过程可视化

现阶段数据挖掘大多都基于相应算法展开，其算法过程不易被使用者直观了解到，所以数据挖掘可视化研究具有一定的研究意义。如将数据挖掘过程可视化处理，可方便用户理解挖掘的整个过程，便于用户实施数据挖掘的操作。

4.2.3 数据挖掘与多库系统的集成

数据库系统、Web 数据库现如今在信息处理系统中成为了主流，数据挖掘系统的理想体系结构是与数据库和数据仓库系统的紧密耦合[3]。

4.2.4 描述语言标准化

研究人员可趋向数据挖掘语言标准化的研究，使数据挖掘语言像 SQL、C++、Java 语言一样标准化、形式化。

4.2.5 复杂数据分析建模方法

大数据时代背景下，数据类型逐渐增多，数据结构独特且逐渐变复杂。为了处理这些复杂和独特数据，需要进一步优化和新增数据分析和建立模型的方法，使后期开展数据挖掘更加容易。

5 结语

信息时代的大爆发，使得各种数据资源迅猛增加，然而数据的增加与数据分析的滞后差值也越来越大，而大多数研究者希望通过科学手段挖掘数据深层价值，所以数据挖掘变成了解决数据分析问题的主流技术，它弥补了传统分析方法的不足，有针对性地对数据进行科学化处理。只有将数据隐藏的有效知识信息及时发现，才能进一步服务于人类发展，数据资源才能真正被利用起来，也才意味着大数据时代的真正到来。

参考文献

[1] 陈艳红. 高校信息系统中的数据挖掘与学生行为预警分析研究 [D] . 湖北民族大学，2019.

[2] 刘喆. 基于潜在语义的 K-means++算法改进及搜索应用的研究与实现 [D] . 南京邮电大学，2019.

[3] 陶翠霞. 浅谈数据挖掘及其发展状况 [J] . 科技信息，2008，(4)：72.

(上接第 114 页)

要确定授权管理制度，这样才能进一步明确身份认证功能以及访问行为记录功能，通过加强访问控制确保访问权限范围，并利用网关工具对数据进行全面管理。

4.3.3 数据安全

商业银行可以对相关的数据进行加密处理，达到提升网络加密技术安全性的目的。首先，对相关数据档案进行加密处理，对于所有电子数据档案要加强相关的采集、存储、传递、使用、复制、传输以及销毁等工作，可以采用异地备份的方式，对相关的数据进行保护处理；此外，商业银行也可以利用合同的方式对数据进行加密处理，不得输入涉及安全性能的数据。

4.3.4 外包技术

信息技术外包工作主要包括外包资格审查制度，在开展相关外包业务时，一定要加强对外包信息技术的监管以及控制工作，为商业银行后续业务的可持续发展打下坚实基础。

5 结语

信息安全系统对商业银行安全运行有着重要的支持以及辅助作用，是不可替代的时代产物。我国商业银行在发展至今，已经具备完整的信息安全体系，但是由于银行业务较为复杂，部分基层人员仍然不具备足够的专业能力以及安全意识，在技术方面仍然存在一定的局限性。加之由于存在不可抗的外界因素，商业银行的信息安全管理体系还需进一步完善。

参考文献

[1] 王逸秋. 刍议商业银行信息系统风险管理方法与对策 [J] . 现代商业，2013，(6)：91.

[2] 李丹. 对中国银行业 “十二五” 信息科技发展规划的思考 [J] . 中国金融电脑，2011，(3)：9-15.