●化柏林 (中国科学技术信息研究所，北京 100038)

数据挖掘与知识发现关系探析

摘 要：以数据挖掘与知识发现的分类为切入点，详细探讨数据挖掘与知识发现的关系。总结出关于数据挖掘与知识发现的关系问题有三种观点，即数据挖掘就是知识发现，数据挖掘是知识发现的一个步骤，数据挖掘与知识发现是完全不同的两个概念。三种观点各有道理，取决于研究者的研究背景、研究范畴与目标。最后对数据挖掘与知识发现的发展趋势进行探讨。

关键词：数据挖掘；知识发现；相关性研究；规则

Abstract: Proceeding from the classification of data mining and knowledge discovery, this paper discusses the relationship between Data Mining (DM) and Knowledge Discovery (KD) in depth. There are 3 viewpoints about the relationship between DM and KD, that is, DM is KD, DM is a step of KD, and DM is entirely different from KD. The 3 viewpoints are all reasonable. It depends on the research background, research scope and goal of the researchers. Finally, the paper discusses the development trend of DM and KD.

Keywords: data mining; knowledge discovery; relativity research; rules

与数据挖掘与知识发现相近或相关的概念有很多，包括数据挖掘 (Data Mining)、知识发现 (Knowledge Discovery)、数据库知识发现 (Knowledge Discovery in Databases, KDD,)、知识挖掘 (Knowledge Mining)、知识元挖掘、知识抽取 (Knowledge Extraction)、信息抽取 (Information Extraction)、信息发现 (Information Discovery)、智能数据分析 (Intelligent Data Analysis)、探索式数据分析 (Exploratory Data Analysis)、信息收获 (Information Harvesting)、数据考古 (Data Archeology) 等。这些概念看似相近，实则不同。在这些概念中，认可度最高的当属数据挖掘与知识发现。数据挖掘与知识发现是否相同，或者存在某种内在的联系，这一关系问题值得深入探讨。

1 数据挖掘

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的但又是潜在有用的信息和知识的过程。现有的数据挖掘技术分为 5 类，即预测模型化、聚类、数据归纳、依赖模型化以及发现变化和偏差。按照数据类型，数据挖掘分为数值挖掘，文本挖掘、Web挖掘、图形图像挖掘、音频挖掘、视频挖掘[1]。数值挖掘是数据挖掘的主流，已有成熟的商业应用，如零售业、金融保险业等。文本挖掘由构造文本集合、文本分析和特征修剪 3个步骤构成，文本挖掘的过程可以细化为从文献源到初始文献集的子过程，从初始文献集到中间词集的子过程，从中间词集到关联词集的子过程[2]，文本挖掘有一定难度。Web挖掘指在人为构造的万维网中挖掘有趣的、潜在的、有用的模式及隐藏的信息的过程。Web挖掘分为 Web内容挖掘、结构挖掘、用法挖掘[3]。挖掘一般在内容分析的基础上进行。图形图像、音频、视频的内容分析还很难，所以挖掘就更难了。而在这些不同类型的数据中，人们最习惯用文本型数据来描述知识，所以从文本中挖掘知识就成为挖掘的重点。

数据挖掘按照数据库模型分为基于关系数据库的数据挖掘、基于面向对象数据库的数据挖掘等。伴随数据库技术的发展，多媒体数据库的数据挖掘、时态数据库的数据挖掘、空间数据库的数据挖掘等也引起了许多人的关注[1]。数据挖掘的任务主要是关联分析、聚类分析、分类、预测、时序模式和偏差分析等。数据挖掘的方法分为统计方法、机器学习方法、神经网络方法等类别。在众多的数据挖掘方法中，关联规则一直是众多学者的研究热点。关联规则的挖掘包括确定性关联规则的挖掘、不确定关联规则的挖掘[4]、量化关联规则的挖掘[5]、增量式关联规则的挖掘[6]、模糊关联规则的挖掘[7]、广义关联规则的挖掘[8]等。关联规则挖掘的目的是在数据库中发现各数据项之间的关联关系。

如果认定数据挖掘是从数据中挖掘的话，那文本挖掘就是从文本数据中挖掘。当数据挖掘特指数值型数据时，那么文本挖掘与数据挖掘是并列的关系，只不过文本挖掘从技术上更难一些。数据值型数据挖掘只考虑数据之间的关系，对某个数据元素本身不需要深入分析。而文本挖掘

除了分析数据之间的关系外，还要对每个数据元素本身进行深入分析，包括形态分析、结构分析以及语义分析。当数据挖掘泛指各种类型数据时，文本挖掘是数据挖掘的一个分支，是它的一个下位类。文本挖掘除了继承数据挖掘的通用方法外，有自己的处理过程与分析方法。

2 知识发现

海量数据与知识贫乏导致了数据挖掘和知识发现研究的出现。知识发现是从数据中识别出有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程[9]。现在的知识发现主要有两大分支，分别为数据库知识发现与基于文献的知识发现。

知识发现的方法分为统计方法、机器学习方法与神经计算方法。统计方法除了回归分析（多元回归、自回归等）、判别分析（贝叶斯判别、费舍尔判别、非参数判别等）、聚类分析（系统聚类、动态聚类等）及探索性分析（主成分分析、相关分析）等方法以外，还包括模糊集方法，支持向量机方法、粗糙集等方法。常用的机器学习方法包括规则归纳、决策树、范例推理、遗传算法等。常用的神经计算方法包括自组织映射网络、反传网络等[10]。

数据库知识发现主要针对结构化数据，统计领域研究的较多，数据库管理系统主要是关系型数据库。基于文献的知识发现主要针对非结构化数据，按照文献的相关性分为基于相关文献的知识发现、基于非相关文献的知识发现和基于全文献的知识发现[11]，但无论是基于何种文献集合的知识发现，针对全文数据是其共同特征。

既然是非相关文献的知识发现，那么这些文献是不能聚类到一起，或者是文献集合聚类之后，不同类的文献之间进行知识发现。知识发现可以由人来进行，也可以由计算机来进行。如果人来发现，那么由谁来阅读非相关的文献呢（往往是不同领域的文献）？找不到这样的专家，那么该由计算机来完成。由计算机直接从非相关文献中发现新知识是非常困难的，应由计算机首先把文献中的知识元抽取出来，构成知识库，然后再在知识库里进行发现。因此知识抽取是知识发现的前提。

基于文献的知识发现在图书情报领域研究较多。非相关文献的知识发现属于知识发现的一类，从分类学的思想来看，非相关文献的知识发现属于知识发现的下位类，而与非相关知识发现的同位类应该还有相关文献的知识发现。因此，根据数据挖掘的定义，知识发现应该是处理之前看起来并不相关的，也就是说，这种相关性是新颖的。正如发现了啤酒和尿布的关系属于知识发现，而面包和牛奶的关系不能属于知识发现。

3 关联规则挖掘与非相关文献知识发现的差异性

数据挖掘中，关联规则用得最多，而知识发现中，基于非相关文献的知识发现受到了越来越多的关注。

数据挖掘中的关联规则与非相关文献的知识发现的本质也是相同的。正如啤酒和尿布没有直接关系，这类似于两篇不同领域的非相关文献。买啤酒的人同时也往往买尿布是新发现的结果，但处理的原始数据并非如此，把它拆开来看，某些人（集合 A1）买啤酒（B）是一条信息（Ⅰ），某些人（集合 A2）买尿布（C）也是信息（Ⅱ），这两条信息中有共同的概念实体，A1和 A2这两个集合所形成的交集，啤酒和尿布没有直接关系，但是这两个概念通过概念 A建立了关系，A就是 A1与 A2的交集，Ⅰ中包含概念 A与 B，Ⅱ中包含概念 A与 C，两条信息中包括 3个概念，就概念数量及概念之间的关系来讲，关联规则的数据挖掘与非相关文献的知识发现存在着共同之处。所不同的是，关联规则中不同的信息元之间往往使用相同的概念关系（如例中的购买）。

非相关文献的知识发现，在两个知识元之间有个共同的主题[12]，每个知识元一般有两个及以上的概念，用一个概念关系把两个概念联系起来形成一个知识元，如知识元 K1：材料 A中可以提取成分 B；知识元 K2：成分 B可以治疗疾病 C，而在另一个知识元中用另外一个概念关系把两个概念联系起来形成另一个知识元，每个知识元中蕴涵着两个概念，但两个知识元却只有 3个概念。

无论是基于数据挖掘的关联规则还是基于非相关文献的知识发现，两个（信息或知识）元之间必然存在共同的概念或概念关系。这种共同关系是挖掘与发现的前提，共同概念的识别是比较容易的。一般来讲，都是基于形式的分析，只是个别情况需要进行语义验证。

4 数据挖掘与知识发现的关系

关于数据挖掘与知识发现的关系主要有 3种观点：数据挖掘就是知识发现；数据挖掘是知识发现的一个步骤[1,9]；数据挖掘与知识发现是完全不同的两个概念。

第一种观点：数据挖掘就是知识发现。数据挖掘是从数据中挖掘，知识发现并不是从知识中发现，而是发现知识。知识是从数据中发现的，是经过挖掘发现的。数据挖掘是从源头入手，知识发现视目标而论。前者强调过程，后者强调结果，应该是一个概念的两种表述，强调重点有所不同而已。如果认为知识发现是从知识中发现的话，按照信息链或信息转化理论，发现的结果应该是智能或情报，而不是知识。从知识中发现新的知识，叫知识创新，所经过的过程叫知识推理。所以说知识发现的概念是不确

切的，准确的说法是数据库知识发现，或者说知识发现是数据库知识发现的简化说法。

第二种观点：数据挖掘是知识发现的一个步骤，数据库知识发现指从数据中获取有用知识的整个过程。数据挖掘是从数据中抽取模式的具体算法的应用。KDD过程除了数据挖掘之外，还有数据预处理、数据筛选、数据清洗、已有匹配知识的吸收、结果的解释与评估，以确保从数据中抽取的知识是有用的[9]。数据挖掘首先从数据集中选出目标数据，经过滤重、去噪、数据清洗等预处理后，进行降维，减少后续过程中需要考虑的特征或变量个数，然后从变换后的数据中挖掘出模式（Pattern），最后经过解释与评价以后变成有用的知识[9]。数据挖掘方法的盲目应用会导致发现一些无意义或者无效的模式。

第三种观点：数据挖掘与知识发现是完全不同的两个概念。数据挖掘主要针对结构化数据，其数据项是不可分割的，符合一范式（1NF）；而知识发现的处理对象是半结构化与非结构化的知识，数据项可以进一步分割，不符合1NF。数据挖掘主要运用回归分析、主成分分析、多元分析、关联规则、支持向量机、模糊集等方法[1]，走统计与规则的技术路线；而知识发现主要是通过神经网络、遗传算法、决策树、范例推理、贝叶斯信念网络等方法[10]，走归纳与演绎的推理过程。数据挖掘的结果往往是精确的、定量的（尽管有置信度这样一个指标）；知识发现的结果往往是模糊的、定性的。数据挖掘主要应用于统计、数据分析等领域；而知识发现主要应用于人工智能领域。关于数据挖掘与知识发现的对比分析如表1所示。

表1 数据挖掘与知识发现对比分析

	数据挖掘	知识发现
处理对象	数据	知识
数据项	不可分割	可以分割
数据特点	结构化	半结构化、非结构化
技术路线	统计、规则	推理、归纳与演绎
主要方法	回归分析、主成分分析、多元分析、关联规则、支持向量机、模糊集等	神经网络、遗传算法、决策树、范例推理、贝叶斯信念网络等
结果	精确的、定量的	模糊的、定性的
应用领域	统计、数据分析	人工智能

5 知识挖掘与文本挖掘

如果认同数据挖掘是从数据中挖掘，最终发现新知识的话，那么知识挖掘应该是从知识中挖掘，从知识中挖掘什么呢？所以知识挖掘不是从知识中挖掘某某东西，而是从某某东西中挖掘知识。从这个角度讲，知识挖掘与知识发现的本质是一回事。

知识元也就是知识点，所以知识元挖掘与知识发现也是一回事，只不过更加突出知识组织的粒度以知识元为单位，即一条一条的知识。温有奎教授的系列论文[13]与《知识元挖掘》[14]一书中，把知识抽取的过程与知识发现的过程合在一起，统称为知识元挖掘。先从文本数据中把显性表示的知识元抽取出来，形成以知识元为单位组织起来的知识库，然后试图从知识库中找出新知识。下面从非相关文献的知识发现的经典案例分析知识抽取与知识发现的关系。“文献x中论述到，材料A可以提取成分B，文献y中论述到成分B可以治疗疾病C”。

可以得出：

R1=材料A可以提取成分B

R2=成分B可以治疗疾病C

那么根据R1和R2可以推断出：

R3=材料A可以治疗疾病C

根据R1和R2推出R3是知识发现的过程，R1和R2的来源问题属于知识抽取的过程。R1和R2是从文献中抽取的知识，是用文本来表述的。绝大多数知识都是以文本形式展现的，因此知识挖掘主要指文本挖掘。同样，知识挖掘是指从文本中挖掘知识，文本挖掘强调处理对象，知识挖掘强调处理结果。

6 结束语

无论哪一种观点都各有其道理，取决于研究背景、研究范畴与研究目标。总体上来讲，更多的学者认同数据挖掘是知识发现的一个步骤。

无论是数据挖掘还是知识发现，目前还没有形成完整的适合中文信息处理的挖掘理论与技术框架。中文文本的特征提取与表示大多数采用“词袋”法[15]。基于词的层面进行信息处理还有很大的问题，因为词很难表达完整的规则知识。大规模文本处理不可能绕过自然语言处理，自然语言处理的过程较为复杂，从分析层面来讲，包括语形分析、语法分析、语义分析和语用分析。基于文本数据的知识发现的分析处理过程不仅仅停留在词法分析、语法分析层面上，还要深入到语义分析甚至语用分析。

数据挖掘更多地集成异构数据，包括数值型数据、文本型数据、图形图像数据、音频数据、视频数据，除此之外，还将深入挖掘空间数据、生物数据等，大型异构异质数据之间的集成与融合将会为数据挖掘提供更多的信息与知识。数据挖掘除关键字匹配技术、统计学技术以外，更多地利用自然语言理解技术实现概念挖掘、结构挖掘、外形挖掘等高级挖掘。如何将模糊技术、免疫进化、粗糙集、支撑向量机理论和技术运用于数据挖掘，实现柔性的数据挖掘是以后的研究方向。挖掘的结果将更倾向于可视化、可理解化。知识发现的数据集合会由单一的数据库向集成的跨库跨系统发展[16]，会由非相关文献的知识发现

向着全文献的知识发现集合发展。知识发现的方法更多地运用人工智能理论方法，更多地研究人类思维与计算机思维的差异与共性，更多地运用逻辑推理、神经计算、遗传算法等理论，将进一步借助自然语言处理技术，从语义的角度深入内容分析，研究方法与研究结果将更多地体现出多学科特性。□

参考文献

[1] Han Jiawei, Kamber M. 数据挖掘概念与技术 [M]. 范明，孟小峰，译. 北京：机械工业出版社，2007

[2] 张云秋，冷伏海. 基于非相关文献知识发现中的文本挖掘研究 [J]. 情报理论与实践，2007，30（2）：194-197

[3] 高岩，胡静涛. Web数据挖掘的原理、方法及用途 [J]. 现代图书情报技术，2002（3）：51-53

[4] 吴涛. 基于概念的不确定性关联规则挖掘 [J]. 现代计算机：专业版，2007（5）：11-13

[5] 佟强，周园春，吴开超，等. 一种量化关联规则挖掘算法 [J]. 计算机工程，2007（10）：34-36

[6] 王云岚，李增智，屈科文. 基于候选项集个数上阶的增量式关联规则更新算法 [J]. 电子学报，2004，32（5）：731-734

[7] 曾庆花，王文国. 一种改进的模糊关联算法及其在DS中的应用 [J]. 计算机技术与发展，2007，17（7）：236-239

[8] 李天瑞，杨宁，马骏. 广义关联规则及其挖掘算法 [J]. 西南交通大学学报，2004，39（1）：86-89

[9] Fayyad U, Shapiro G P, Smyth P. From data mining to knowledge discovery in databases [EB/OL]. [2003-06-22]. http://www.kdnuggets.com/gpspubs/imag-kdd-overview-1996-Fayyad.pdf

[10] 史忠植. 知识发现 [M]. 北京：清华大学出版社，2002

[11] 张树良，冷伏海. 基于文献的知识发现的应用进展研究 [J]. 情报学报，2006，25（6）：700-712

[12] 安新颖，冷伏海. 基于非相关文献的知识发现原理研究 [J]. 情报学报，2006，25（1）：87-93

[13] 温有奎，温浩，徐端颐，等. 基于知识元的文本知识标引 [J]. 情报学报，2006，25（3）：282-288

[14] 温有奎，徐国华，赖伯年，等. 知识元挖掘 [M]. 西安：西安电子科技大学出版社，2005

[15] 谌志群，张国煊. 文本挖掘与中文文本挖掘模型研究 [J]. 情报科学，2007，25（7）：1046-1051

[16] 黄晓斌，邓爱贞. 现代信息管理的深化——数据挖掘和知识发现的发展趋势 [J]. 现代图书情报技术，2003（4）：1-3，16

作者简介：化柏林，男，1977年生，硕士，助理研究员。
研究方向：自然语言处理。

收稿日期：2008-03-27

（上接第491页）

如何推动隐性知识的显性化既是一个系统的社会问题，也是复杂的技术难题，有待深入探讨。

3.3 知识创新

知识资源建设和知识提供保障的最终目的都是为了支持知识创新。知识创新过程是人能动地认识世界和改造世界的过程，产出具有相当大的不确定性。因此，笔者认为不存在绝对的保障，用“支持”一词似乎更确切。

虽然所有知识服务主体都努力将业务向知识创新层拓展，渴望成为社会知识创新体系的核心。但笔者认为，“授人以鱼，不如授人以渔”，最好的保障不是提供创新成果，而是培育各社会主体的持续创新能力。从营造组织知识创新微生态，到形成整个国家和社会的良好知识生态，才是最全面的知识创新保障。这是一个有待开发的广阔空间，似乎谁都可以介入，但谁都没成为引领和主导者，这里没有垄断，甚至不存在竞争。

综上所述，知识资源应当由国家投资建设，政府机构和图书情报机构是提供保障的主体，这是当前国家知识保障工作的重心。知识提供保障必须借助国家和市场两种力量，即以法律和制度确定政府机构和高等院校的知识提供义务，以市场作为利益分享机制，调动各方面的主动性和积极性，实现需求与供给的均衡；知识创新是社会的共同责任，国家知识创新保障不可能依靠少数主体实现，但是，知识服务机构作为营造社会和谐知识生态的主导力量将发挥举足轻重的作用。□

参考文献

[1] 周城雄. 隐性知识与显性知识的概念辨析 [J]. 情报理论与实践，2004（2）：127-129

[2] 唐雷工作室. 隐性知识和显性知识 [EB/OL]. [2007-11-05]. http://www.tlstudio.net/post/content/2006-5/content_261.html

[3] 深蓝雨. 情报商品化的发展过程 [EB/OL]. [2007-11-10]. http://hi.baidu.com/vhion/blog/item/e1be7209776f27cf3bc763b7.html

[4] 包昌火. 情报研究的产生和发展 [EB/OL]. [2007-11-05]. http://www.defence.org.cn/?article-13-64141.html

[5] 陈搏，等. 知识资源池：知识创新和共享的宏观机制模型 [J]. 科学学研究，2006（S）：274-279

[6] 中国政府网. 中华人民共和国政府信息公开条例 [EB/OL]. [2007-11-08]. http://www.gov.cn/zwgk/2007-04/24/content_592937.htm

作者简介：段宇锋，男，博士，副教授。

收稿日期：2008-01-14