Data and text mining and its application in research and development decision-making

数据与文本挖掘及其在研发决策中的应用

郝占刚 著

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

数据与文本挖掘及其在研发决策中的应用

郝占刚 著

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

图书在版编目（CIP）数据

数据与文本挖掘及其在研发决策中的应用/郝占刚著.
—北京：经济管理出版社，2011.12
ISBN 978-7-5096-0855-5

Ⅰ.①数… Ⅱ.①郝… Ⅲ.①数据采集 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2011）第 259461 号

出版发行：经济管理出版社
北京市海淀区北蜂窝 8 号中雅大厦 11 层
电话：(010)51915602 邮编：100038

印刷：北京银祥印刷厂 经销：新华书店

组稿编辑：申桂萍 责任编辑：杨国强 高 蕙
责任印制：杨国强 责任校对：曹 平

720mm×1000mm/16 10.5 印张 136 千字
2011 年 12 月第 1 版 2011 年 12 月第 1 次印刷

定价：32.00 元

书号：ISBN 978-7-5096-0855-5

前 言

随着数据库技术的不断发展及数据库管理系统的广泛应用，数据库中存储的数据量急剧增大，在这些大量的数据背后隐藏着许多重要的信息，如果能把这些信息从数据库中抽取出来，将为数据的所有者创造出很多潜在的利润和价值，而这种从海量数据库中挖掘信息的技术，就称为数据挖掘。数据挖掘技术可以解决“数据爆炸但知识贫乏”的现象。而存储信息使用最多的是文本，所以文本挖掘被认为比数据挖掘具有更高的商业潜力。当数据挖掘的对象完全由文本这种数据类型组成时，这个过程就称为文本挖掘。事实上，有研究表明公司信息有80%包含在文本文档中。因此，数据挖掘和文本挖掘成了目前学术界和应用领域重点研究的问题。

进化算法是以进化论为思想基础，通过模拟生物进化过程与机制的求解问题的自组织、自适应的人工智能技术。本书主要研究进化算法中的遗传算法和社会演化算法在数据与文本挖掘中的应用。

遗传算法是借鉴生物界自然选择和遗传机制的随机搜索优化算法，作为一种有效的全局并行优化搜索工具，在数据挖掘领域得到了广泛的应用，是数据挖掘的主要算法之一。基于遗传算法的特点，遗传算法在数据挖掘的三大研究领域：数据收集和预处理、挖掘、评价和知识呈现方面得到了广泛的应用，取得了良好的效果。社会演化算法的思想基础是库恩的范式转换理论，其寻优机制是基于范式的确立与更新以及认知主体对范式进行学习的一系列智能认知行为。该算法目前主要应用于解决组合优化问题，用于数据与文

本挖掘还很少有人研究。

本书运用遗传算法、社会演化算法等进化算法，结合k-均值算法、k-medoids算法、神经网络方法、模式聚合方法、潜在语义索引等方法对数据和文本挖掘中的特征降维问题、分类问题、聚类问题等进行了研究，提出了一些新的高效的算法，并将这些算法运用到企业产品研发决策中去，提高产品研发的效率。本书的主要内容如下：

第一，提出一种基于遗传算法和k-medoids算法的新的聚类方法。采用遗传算法进行聚类，或时间成本太高，或效果不佳。因此，本书提出将k-medoids算法嵌入遗传算法中，形成一个新的聚类算法。该方法既可以很好地解决局部最优的问题，又可以很好地解决孤立点的问题，同时用于和k-medoids算法相结合，可以加快遗传算法的收敛速度，节约时间成本。

第二，采用遗传算法和模式聚合进行文本特征降维。文本向量一般都具有非常高的维数，几千维或上万维，这么高的维数使得文本挖掘的效率非常低，本书提出了遗传算法和模式聚合相结合的降维算法。模式聚合可以有效降低文本特征的维数，使特征从几千维降为几百维，并在此基础上采用遗传算法继续降维。

第三，采用遗传算法和潜在语义索引进行文本特征降维。潜在语义索引通过奇异值分解可以有效降低向量空间的维数，并在此基础上采用遗传算法继续降维。

第四，采用社会演化算法进行聚类。K均值聚类算法通常只能以局部最优结束，很难找到全局最优。本书提出一种基于社会演化算法和k-均值算法相结合的聚类新方法，在该方法中提出了认知主体在聚类中对范式学习的新的方式。

第五，采用混沌社会演化算法进行聚类。在认知主体对范式的背叛中采用混沌变异算子。实验证明该方法不但能提高聚类的效率，而且能提高聚类的精度。

第六，将 k-medoids 算法和改进的遗传算法、改进的社会演化算法相结合，解决文本聚类及孤立点问题。将 k-medoids 算法嵌入遗传算法中，将其聚类结果作为遗传算法的初始种群，并进而在每一代都采用其进行优化，从而缩减遗传算法进化时间，提高进化效率和质量。将 k-medoids 算法嵌入社会算法中，将其作为智能认知主体的认知算法。

第七，文本挖掘在产品研发决策的应用研究。产品研发是一个知识融合、传递和共享的过程，不同知识之间的嫁接、变异、融合，形成了新的技术，从而诞生新的产品。因此，产品研发的关键在于对知识的利用。而文本挖掘方法就是一个非常好的将文本数据转化为文本知识的方法。本书在文本数据搜集的基础上，采用文本挖掘方法对其进行处理，获得产品研发所需的文本知识，进而构建产品研发文本知识地图。该知识地图可以帮助企业进行产品研发决策，提高产品研发的质量和效率。

郝占刚

2011 年 11 月

目 录

第一章 绪 论

本章首先介绍选题的研究背景和意义，然后对数据挖掘和文本挖掘的主要概念、过程、技术等进行阐述，接下来重点介绍了遗传算法的基本概念及其在数据挖掘和文本挖掘中的应用以及相关的研究概况，最后介绍了本研究的主要工作。

第一节 本书的研究背景和意义

随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。

数据挖掘是从大量原始数据中发掘出隐含的、有用的但尚未被发现的信息和知识（如规则、规律、模式、约束等），目的是帮助决策者寻找数据间潜在的关联，发现被忽略的要素，而这些信息对预测趋势和决策行为是十分有用的。[1,5,17,19,23,29] 数据挖掘技术能从数据库中自动分析数据，进行归纳性推理，从中发掘出潜在的模

式；或者产生联想，建立新的业务模型，帮助决策者做出正确的决策。数据挖掘表明，知识就隐藏在日常积累下来的大量数据之中，但是发现这些知识是需要借助有效的方法和工具的。

遗传算法作为一种有效的全局并行优化搜索工具，在数据挖掘领域得到了广泛的应用，是数据挖掘的主要算法之一。遗传算法是一类借鉴生物界自然选择和遗传机制的随机搜索优化算法，其主要特点有：它处理的对象不是参数本身，而是对参数集进行编码后得到的个体，因此不仅可以对传统的目标函数优化求解，而且可以处理诸如矩阵、树和图等结构形式的对象；它采用同时处理群体中多个个体的方法，即同时对搜索空间中的多个解进行评估，降低了陷入局部最优解的风险。另外，遗传算法基本上不需要有关搜索空间的知识或者其他辅助信息，仅仅靠适应度函数值来评估个体，适应度函数不受连续可微等条件的约束。基于遗传算法的特点，遗传算法在数据挖掘的三大研究领域：数据收集和预处理、挖掘、评价和知识呈现方面得到了广泛的应用，取得了良好的效果。

社会演化算法是建立在社会认知模型（Social Cognitive Model，SCM）之上的一种算法。SCM 的参照系统是人类社会，所以该模型的组织结构与人类社会本身有许多相似之处。与人类社会由大量个体构成相似，SCM 由许多认知主体（Cognitive Agent）构成。[141] 一个认知主体就是一个具有一定简单的推理、决策等认知能力的人工系统。每一个认知主体都有独立的认知能力，经过一系列认知行为后可以得到一个局部最优解。社会演化算法的思想基础是库恩的范式转换理论，其寻优机制是基于范式的确立与更新以及认知主体对范式进行学习的一系列智能认知行为。社会演化算法是一种优良的寻优算法，能够解决数据挖掘和文本挖掘方面的问题。

数据挖掘的大部分研究主要针对的是结构化的数据，如关系的、事物的和数据仓库的数据。然而当今是一个信息产生、传播速度很快，信息交流量日益增加的时代。由于电子形式的信息量的飞

速增长，如电子出版物、电子邮件、CD-ROM 和万维网（它也可以被视为一个巨大的、互联的动态文本数据库）等，文本数据库得到迅速的发展。文本数据库中存储最多的数据是所谓的非结构化数据。文本挖掘由两步组成：预处理过程和发现过程。文本挖掘是从大量文本中发现新的知识、进行文本处理的新的研究领域。

数据挖掘方法的提出，让人们认识到数据的真正价值，即蕴涵在数据中的信息和知识。数据挖掘是目前国际数据库和信息决策领域最前沿的研究方向之一，已经引起了学术界和工业界的广泛关注。

综上所述，数据挖掘和文本挖掘是当前的一个重要领域，目前将遗传算法理论应用到数据挖掘和文本挖掘中，已提出了一些方法，还存在很多值得研究的内容；而社会演化算法还未见用于数据挖掘和文本挖掘的研究当中，具有很大的研究价值。故基于遗传算法和社会演化算法的数据及文本挖掘具有较大的理论意义和实用价值。

第二节 数据挖掘与文本挖掘概述

一、数据挖掘概述

1. 数据挖掘的定义及发展

随着数据库技术的不断发展及数据库管理系统的广泛应用，数据库中存储的数据量急剧增大，在这些大量的数据背后隐藏着许多重要的信息，如果能把这些信息从数据库中抽取出来，将为数据的

所有者创造出很多潜在的利润和价值，而这种从海量数据库中挖掘信息的技术，就称为数据挖掘。[2]

数据挖掘是从大型数据库或数据仓库中发现并提取隐藏在其中的信息的一种新技术，目的是帮助决策者寻找数据间潜在的关联，发现被忽略的要素，而这些信息对预测趋势和决策行为也许是十分有用的。数据挖掘技术涉及数据库、人工智能、机器学习和统计分析等多种技术。数据挖掘技术能从大型数据库或数据仓库中自动分析数据，进行归纳性推理，从中发掘出潜在的模式；或者产生联想，建立新的业务模型，帮助决策者调整市场策略，做出正确的决策。[4,18,39,40] 数据挖掘表明：知识就隐藏在日常积累下来的大量数据之中，而仅靠复杂的算法和推理并不能发现知识。

KDD 一词首次出现是在 1989 年 8 月举行的第 11 届国际联合人工智能学术会议上。迄今为止，由美国人工智能协会主办的 KDD 国际研讨会已经召开了八次，规模由原来的专题讨论会发展到国际学术大会，研究重点也逐渐从发现方法转向系统应用，并且注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。其他内容的专题会议也把数据挖掘和知识发现列为议题之一。KDD 成为当前计算机科学界研究的一大热点。[3,41]

此外，数据库、人工智能、信息处理、知识工程等领域的国际学术刊物也纷纷开辟了 KDD 专题或专刊。IEEE 的 Knowledge and Data Engineering 会刊在 1993 年出版了 KDD 技术专刊，代表了当时 KDD 研究的最新成果和动态，较全面地论述了 KDD 系统方法论、发现结果的评价、KDD 系统设计的逻辑方法。不仅如此，在互联网上还有不少 KDD 电子出版物，其中以半月刊“Knowledge Discovery Nuggets”最为权威，还可以下载各种各样的数据挖掘工具软件和典型的样本数据仓库，供人们测试和评价。[6,7]

在数据挖掘技术日益发展的同时，许多数据挖掘的商业软件工具也逐渐问世。数据挖掘工具主要有两类：特定领域的数据挖掘工

具和通用的数据挖掘工具。特定领域的数据挖掘工具针对某个特定领域的问题提供解决方案。在设计算法时，充分考虑到数据、需求的特殊性，并作了优化。例如，IBM 公司的 Advanced Scout 系统针对 NBA 的数据，帮助教练优化战术组合；芬兰赫尔辛基大学计算机科学系开发的 TASA，帮助预测网络通信中的警报。特定领域的数据挖掘工具针对性比较强，只能用于一种应用。也正因为针对性强，往往采用特殊的算法，可以处理特殊的数据，实现特殊的目的，发现的知识可靠度也比较高。通用的数据挖掘工具不区分具体数据的含义，采用通用的挖掘算法，处理常见的数据类型。例如，IBM 公司 Almaden 研究中心开发的 QUEST 系统和 SGI 公司开发的 MineSet 系统。通用的数据挖掘工具可以做多种模式的挖掘，挖掘什么、用什么来挖掘都由用户根据自己的应用来选择。

随着 KDD 研究逐步走向深入，人们越来越清楚地认识到，KDD 的研究主要有三个技术支柱，即数据库、人工智能和数理统计。目前数据库界除了关注万维网数据库、分布式数据库、面向对象数据库、多媒体数据库、查询优化和并行计算等技术外，已经开始反思，数据库最实质的应用是否仅仅是查询。理论根基最深的关系数据库最本质的技术进步点，就是数据存放和数据使用之间的相互分离。然而，人们越来越清楚地发现，“查询是数据库的奴隶，发现才是数据库的主人”。[8,20]

由于数据库文化的迅速普及，用数据库作为知识源具有坚实的基础；另外，对于一个感兴趣的特定领域——客观世界，先用数据库技术将其形式化并组织起来，就会大大提高知识获取起点，以后从中发掘或发现的所有知识都是针对该数据库而言的。因此，在需求的驱动下，很多数据库学者转向对数据仓库和数据挖掘的研究、从对演绎数据库的研究转向对归纳数据库的研究。

专家系统曾经是人工智能研究工作者的骄傲。专家系统实质上是一个问题求解系统，目前的主要理论工具是基于谓词演算的机器

定理证明技术——二阶演绎系统。领域专家长期以来面向一个特定领域的经验世界，通过人脑的思维活动积累了大量有用信息。

在研制一个专家系统时，知识工程师首先要从领域专家那里获取知识，这一过程实质上是归纳过程，是非常复杂的个人到个人之间的交互过程，有很强的个性和随机性。因此，知识获取成为专家系统研究中公认的“瓶颈”问题。其次，知识工程师在整理表达从领域专家那里获得的知识时，用 If–Then 等类的规则表达，约束性太大；用常规数理逻辑来表达社会现象和人的思维活动局限性太大，也太困难，勉强抽象出来的规则有很强的工艺色彩，差异性极大。最后，即使某个领域的知识通过一定手段获取并表达了，但这样做成的专家系统对常识和百科知识出奇的贫乏，而人类专家的知识是以拥有大量常识为基础的。人工智能学家 Feigenbaum 估计，一般人拥有的常识存入计算机大约有 100 万条事实和抽象经验法则，离开常识的专家系统有时会比傻子还傻。例如战场指挥员会根据“在某地发现一只刚死的波斯猫”的情报很快断定敌高级指挥所的位置，而再好的军事专家系统也难以顾全到如此的信息。

以上这三大难题大大限制了专家系统的应用，使得专家系统目前还停留在构造诸如发动机故障诊断一类的水平上。人工智能学者开始着手基于案例的推理，尤其是从事机器学习的科学家们，不再满足自己构造的小样本学习模式的象牙塔，开始正视现实生活中大量的、不完全的、有噪声的、模糊的、随机的大数据样本，也走上了数据挖掘的道路。

数理统计是应用数学中最重要、最活跃的学科之一，它在计算机发明之前就诞生了，迄今已有几百年的发展历史。如今相当强大有效的数理统计方法和工具，已成为信息咨询业的基础。信息时代，咨询业更为发达。然而，数理统计和数据库技术结合得并不算快，数据库查询语言 SQL 中的聚合函数功能极其简单，就是一个证明。咨询业用数据库查询数据还远远不够。一旦人们有了从数据

查询到知识发现、从数据演绎到数据归纳的要求，概率论和数理统计就获得了新的生命力。[9,10,11]

2. 数据挖掘的主要研究内容

数据挖掘的任务就是发现隐藏在数据中的模式，其可以发现的模式一般分为两大类：描述型（Descriptive）模式和预测型（Predictive）模式。[12] 描述型模式是对当前数据中存在的事实做规范描述，刻画当前数据的一般特性；预测型模式则是以时间为关键参数，对于时间序列型数据，根据其历史和当前的值去预测其未来的值。根据模式特征，可将模式大致细分如下：

（1）分类模式（Classification）。分类就是构造一个分类函数（分类模型），把具有某些特征的数据项映射到某个给定的类别上。该过程由两步构成：模型创建和模型使用。模型创建是指通过对训练数据集的学习来建立分类模型；模型使用是指使用分类模型对测试数据和新的数据进行分类。其中的训练数据集是带有类标号的，也就是说在分类之前，要划分的类别是已经确定的。通常分类模型是以分类规则、决策树或数学表达式的形式给出的。

（2）聚类模式（Clustering）。聚类就是将数据项分组成多个类或簇，类之间的数据差别应尽可能大，类内的数据差别应尽可能小，即为“最小化类间的相似性，最大化类内的相似性”原则。与分类模式不同的是，聚类中要划分的类别是未知的，它是一种不依赖于预先定义的类和带类标号的训练数据集的非监督学习（Unsupervised Learning），无须背景知识，其中类的数量由系统按照某种性能指标自动确定。

（3）回归模式（Regression）。回归模式的函数定义与分类模式相似，主要差别在于分类模式采用离散预测值（如类标号），而回归模式采用连续的预测值。在这种观点下，分类和回归都是预测问题。但在数据挖掘业界，大家普遍认为用预测法预测类标号为分

类，预测连续值（如使用回归方法）为预测。[13] 许多问题可以用线性回归解决，对于许多非线性问题可以通过对变量进行变换，从而转换为线性问题来解决。

（4）关联模式（Association）。关联模式是数据项之间存在的关联规则，是在同一事件中出现的不同项之间的相关性，比如顾客在同一次购买活动中所购买的不同商品之间的相关性。

最著名的关联规则挖掘算法是由 Agrawal 等人于 1994 年提出的 Apriori 算法。[14,15] Apriori 算法的基本思想是：统计多种商品在一次购买中共同出现的频数，然后将出现频数多的搭配转换为关联规则。Apriori 算法的核心是：用前一次扫描数据库的结果产生本次扫描的候选项目集，从而提高搜索的效率。之后人们又提出了诸多关联规则挖掘算法，主要工作集中在如何提高项集的生成效率和降低计算代价上。

（5）序列模式（Sequential）。序列模式是描述基于时间或其他序列经常发生的规律或趋势，并对其建模。一个典型的例子就是：在购买 PC 机的顾客当中，70%的人会在半年内购买内存条。序列模式将关联模式和时间序列模式结合起来，重点考虑数据之间在时间维上的关联性。有三个参数的选择对序列模式挖掘的结果影响很大：①序列的持续时间，也就是某个时间序列的有效时间或者是用户选择的一个时间段；②时间折叠窗口，在某段时间内发生的事件可以被看作是同时发生的；③所发现模式的时间间隔。

（6）偏差模式（Deviation）。偏差模式是对差异和极端特例的描述，如聚类外的离群值。大部分数据挖掘方法都将这种差异信息视为噪声而丢弃，然而在一些应用中，罕见的数据可能比正常的数据更有用。[13] 比如信用卡的欺骗检测（Fraud Detection），通过检测一个给定账号与其历史上正常的付费相比，可以付款数额特别大这一异常数据为依据来发现信用卡被欺骗性使用。