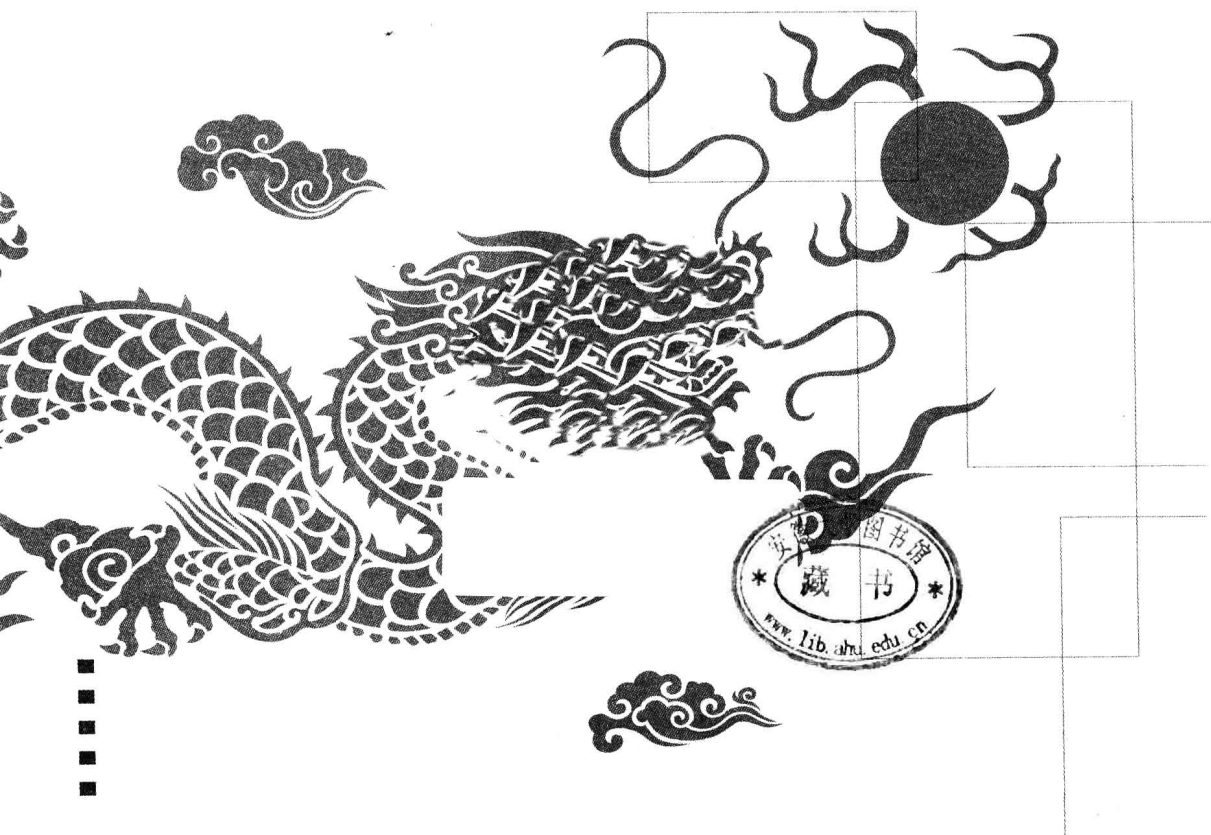
云模型与文本挖掘

代劲 宋娟 胡峰 伍建全 著

人民邮电出版社
POSTS & TELECOM PRESS

代劲 宋娟 胡峰 伍建全 著

云模型与文本挖掘



人民邮电出版社
北京

图书在版编目（CIP）数据

云模型与文本挖掘 / 代劲等著. -- 北京 : 人民邮电出版社, 2013.2
ISBN 978-7-115-30032-4

Ⅰ. ①云… Ⅱ. ①代… Ⅲ. ①人工智能—研究 Ⅳ. ①TP18

中国版本图书馆CIP数据核字(2012)第298709号

内 容 提 要

在当前文本挖掘领域中，传统的数据挖掘方法依然占据着主导地位。然而随着文本挖掘研究的深入，面临着越来越严峻的挑战。这些挑战归根到底是由于自然语言的不确定性造成的。借助不确定性知识研究的重要工具——云模型在定性概念与定量数据间的转换作用，作者将其引入到文本挖掘关键问题研究中，力图降低自然语言中的不确定性知识对文本挖掘性能的影响。在充分利用现有技术的基础上，作者进行了一些大胆的尝试，努力探索出适用于文本挖掘的不确定性人工智能处理方法，用以抛砖引玉，为文本挖掘技术的进一步发展提供一种新的思路与解决方法。

云模型与文本挖掘

◆ 著　　代 劲 宋 娟 胡 峰 伍建全
责任编辑　刘 博

◆ 人民邮电出版社出版发行　　北京市崇文区夕照寺街14号
邮编　100061　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：700×1000　1/16
印张：11.25　　2013年2月第1版
字数：218千字　　2013年2月北京第1次印刷

ISBN 978-7-115-30032-4

定价：48.00元

读者服务热线：(010)67170985　印装质量热线：(010)67129223
反盗版热线：(010)67171154

目　录

第 1 章

绪论

文字是人类文明发展延续的重要产物，记录着人类社会的点滴进步，闪耀着数千年人类智慧之光，是信息的主要载体与知识传播的主要手段之一。文本（Text）是书面语言的文字表现形式，从文学的角度说，通常是具有完整、系统含义（Message）的一个句子或多个句子的组合。一个文本可以是一个句子（Sentence）、一个段落（Paragraph）或者一个篇章（Discourse）。

随着信息技术的快速发展创新，文本的内涵和外延都得到极大的充实，随之而来创建文本、传播文本信息变得十分简单。当前意义上的文本，既包含了纸质的文字信息，也包括了互联网上海量的电子资源。

1.1 文本挖掘的产生背景

随着以互联网为核心的信息高速公路的不断发展与广泛普及，信息技术已经渗透到社会生活中的各个层面，并以前所未有的速度改变着人们的思维、生活及工作方式。网络已经成为拥有海量存储的分布式信息空间，信息量以每 4～6 个月翻一番的速度不断累积。如何在互联网上，从海量异质的信息资源中快速高效地发掘出蕴含其中的具有巨大潜在价值的知识与信息，并且进行合理分类、准确定位，同时筛选掉其中无用或不相关的内容，已经成为知识获取的首要问题。

随着互联网的广泛推广与深入应用，文本正以指数级数量不断翻番。《第 25 次中国互联网络发展状况统计报告》[1]（中国互联网络信息中心（CNNIC），2010 年 1 月）就明确地统计出中国目前的网页总数已达 336 亿个，其年增长率超过 100%。报告中指出，文本信息依然是互联网资源的主要组成部分，比例达到 87.8%。其他一些资源，例如图像、音频与视频所占比例增长较小。根据该报告我们可以看出，虽然互联网上的各种信息资源形式多样且结构复杂，但最重要的信息资源依然是文本。而且，其他形式的信息资源在经过标注技术处理后，均可顺利转换成文本方式。

文本信息的快速增长也使得信息处理技术面临着前所未有的挑战，主要包含以下几个方面的问题：首先，互联网的快速发展，使得文本的不断转载情况变得更为严重，网络上的信息资源存在着大量的重复情况；另外，由于信息资源的结构复杂，冗余严重，目前的信息检索技术还不能有效地搜索到所需信息；最后，网络上的信息资源还面临着严重的信息污染问题，各种垃圾邮件、有害信息还未得到有效的遏制。

这些问题造成信息过剩但知识相对匮乏的现象。与此同时，人们日益增长的各种信息需求已经使得基于传统人工处理的信息抽取、标注、分类、信息过滤及查询越来越不能满足网络化需要。在此基础上，如何利用计算机自动对海量的文本信息进行处理，挖掘出其中有价值的信息，完成海量信息的知识获取过程已经成为一个亟待解决的重大研究课题。文本挖掘技术就是基于这样的背景应运而生并不断发展创新的[2-4]。

1.2 文本挖掘的重要意义

文本挖掘是数据挖掘的一个分支，是以文本作为挖掘对象，从中寻找信息的结构、模型、模式等隐含的具有潜在价值的知识的过程。文本挖掘在信息检索、模式识别、自然语言处理等多个领域均有所涉及。由于文本是信息存储的最主要途径，因此文本挖掘的重要性也日益凸显。

文本挖掘的主要意义在于以下几个方面。

1.2.1 推进信息化建设

劳动工具的改善是人类社会进步开始的标志，随之而来才伴随着生产效率的提高、劳动者的解放。信息文明就是要通过不断发展的 IT 技术，将人们从繁琐的数据采集、统计中解放出来，实现生产和服务运营的智能化。从某种程度来说，这也是对劳动者智力的一次解放。简单来讲，信息化建设的主要核心就是生产和服务流程的自动化、处理方法的智能化。而实现自动化、智能化的主要途径就是通过对信息的深入分析和挖掘，从中发现知识和规则，从而形成对现象及事件运行状态和变化的准确判断。

1.2.2 提高信息利用效率

由于文本数据的表示、存储及输出多样，若不能对其进行有效的转换、分类等操作，其中的丰富信息将不能充分使用。Web 文本搜索引擎技术发展的成功经验表明，文本挖掘不仅能提升 Web 向用户（或终端）信息输出的准确性和效率，还可以大大增强用户对 Web 的信任度。文本挖掘不仅是一项具有较大实用价值的技术，也是组织和管理文本信息的有力手段。通过文本挖掘，可找出隐含在文本信息中的模

式，发现可能忽略的预测信息等。而对企业来讲，文本挖掘就是一种决策支持过程，在人工智能、机器学习、统计学等技术基础上，自动分析处理原有的数据信息，最终获得归纳性的推理、找出潜在的模式并准确预测用户行为。这不仅能帮助决策者调整策略措施，也在一定程度上减少风险，有利于用户做出正确的决策。这也就是文本挖掘的核心问题。

1.2.3 提高人工智能水平

人工智能（Artificial Intelligence，AI）是研究使计算机来模拟人的某些思维过程和智能行为（如学习、推理、思考、规划等）的学科，主要包括计算机实现智能的原理、制造类似于人脑智能的计算机，使计算机能实现更高层次的应用。人工智能涉及计算机科学、心理学、哲学和语言学等学科，可以说几乎涉及自然科学和社会科学的所有学科，其范围已远远超出了计算机科学的范畴。人工智能与思维科学的关系是实践和理论的关系，人工智能是处于思维科学的技术应用层次，是它的一个应用分支。从思维观点看，人工智能不仅限于逻辑思维，要考虑形象思维、灵感思维才能促进人工智能的突破性的发展。

计算机是研究人工智能的主要技术基础，人工智能的发展历史是和计算机科学技术的发展紧密相关的。除了计算机科学以外，人工智能还涉及信息论、控制论、自动化、仿生学、生物学、心理学、数理逻辑、语言学、医学和哲学等多门学科。人工智能学科研究的主要内容包括：知识表示、自动推理和搜索方法、机器学习和知识获取、知识处理系统、自然语言理解、计算机视觉、智能机器人、自动程序设计等方面。

人工智能水平的高低，主要通过判断理解能力、决策思维能力和实施指挥能力进行衡量。其中判断理解能力最为关键。当前，绝大部分信息都是以用自然语言表示的文本形式进行存储。但目前机器还远远不具备对人类自然语言的判断理解能力。所以，将文本翻译成机器可以读取并理解的形式是利用文本信息的前提条件。Web领域、图书信息管理和新闻档案管理领域的成功经验也对此进一步加以验证。经过多年不断的研究与实践，文本挖掘技术已经逐渐与信息推送、信息过滤、搜索引擎等信息处理技术相融合，在图书出版、情报收集、电子阅览、信息服务等大规模使用文本数据库行业取得了丰硕的果实。

1.2.4 保障决策支撑

决策支持系统（Decision Support System，DSS）是辅助决策者通过数据、模型和知识，以人机交互方式进行半结构化或非结构化决策的计算机应用系统。它是管理信息系统（MIS）向更高一级发展而产生的先进信息管理系统。它为决策者提供分析问题、建立模型、模拟决策过程和方案的环境，调用各种信息资源和分析工具，

帮助决策者提高决策水平和质量。

在以往的决策支持系统中，需由专家或程序员建立知识库中的知识和规则，其有效性取决于专家个体的经验和知识水平。而且当数据达到一定规模后，处理及判断这些信息的工作量将会大大超过专家能力。文本数据挖掘的首要任务就是通过相应的挖掘技术发现数据中难以发现的知识或规则，是一个自动获取知识的过程。一方面可以通过查询、联机分析处理等工具直接获取信息提供给决策者；另一方面能找出隐藏在大量数据中的关系、规则和趋势，由系统发现并提供给相应的数据管理专家。

综上所述，文本作为最重要的信息载体，充分利用文本数据挖掘技术对其进行知识获取不仅可以创造巨大的商业价值与社会价值，也是人类社会向信息文明转变过程中必备的信息处理工具。

1.3 不确定性人工智能及其研究方法

1.3.1 不确定性人工智能概述

“不确定性”一词最早出现于1836年詹姆斯·穆勒的著作《政治经济学是否有用》中，随之而来伴随着的关于物质世界确定还是不确定性质的讨论也越演越烈。

以牛顿理论为代表的确定性科学，创造了给世界以精确描绘的方法，将整个宇宙看作是钟表式的动力学系统，处于确定、和谐、有序的运动之中。只要知道初始条件，就可以决定未来的一切。从牛顿到拉普拉斯，再到爱因斯坦，描绘的都是一幅幅完全确定的科学世界图景。确定性论者也并非拒绝一切不确定性，但是他们认为产生不确定的原因是对初始条件的测量误差，或者人类自身认知的局限性和知识的不完备，而并非事物的本来面貌。确定性科学的影响曾经如此强大，以至于在相当长的一段时间内限制了人们认识宇宙的方式和视野，虽然生活在到处都有复杂混乱现象的现实世界里，科学家们看到的却只是钟表式的机械世界，科学的任务只是阐明这架钟表的结构和运行规律，而将不确定性看作是无足轻重的，并将其排除在近代科学的研究对象范围之外。

与物质世界的确定论相对应的，是以概率论、随机性及测不准原理为基础的不确定性理论。麦克斯韦认为这个世界的真正逻辑是概率演算，玻尔兹曼则把随机性观点引入物理学，建立了统计力学。对确定论造成更大冲击的是量子力学的出现。海森堡的测不准原理表明，获得严格精确的初值在原理上是不可能的。测不准原理也从另一方面阐述了这样一个真理：不确定性是客观世界中的一种真实存在，是存在于宇宙间的基本要素，与人类的认知没有任何关系。

客观世界中的绝大部分现象都是不确定的，所谓确定的、规则的现象，只会在

一定的前提和特定的边界条件下发生，只会在局部或者较短的时间内存在。随着不确定性研究的深入，世界的不确定性特征越来越得到学术界的普遍认可，无论是在物理学、数学、生物学等自然科学领域，还是在哲学、经济学、社会学、心理学、认知学等社会科学领域，虽然许多人还在从事着确定性的研究，但已经很难有人对世界的不确定性本质提出实质性的质疑了。越来越多的科学家相信，不确定性是这个世界的魅力所在，只有不确定性本身才是确定的。正是在这样的背景下，混沌科学、复杂性科学和不确定性人工智能才得到了蓬勃发展。

当然，不确定性和确定性并非完全对立，在一定程度上可以相互转化。例如，某一层次的不确定性可能是更高层次上的确定性，种种不确定性中还可能隐藏着某些确定的规律等。人工智能学家的任务，就是寻找并且能够形式化地表示不确定性中的规律性，至少是某种程度的规律性，从而使机器能够模拟人类认识客观世界、认识人类本身的认知过程。

不确定性人工智能就是通过对人类智能的模拟，在人脑的定性分析和机器的定量处理之间建立联系的研究领域。通过对不确定性中的规律性的形式化表示，不确定性人工智能研究能够使机器模拟人类认识客观世界、了解人类自身的认知过程。

1.3.2 不确定性人工智能的主要研究内容

不确定性人工智能主要研究内容可以从两个方面展开：存在论与认识论。存在论主要站在客观物质世界规律性角度，而认识论则从人类主观的认知出发，揭示人类认知的不确定性以及与之对应的自然语言的不确定性。

混沌、分形和复杂网络是当前基于存在论的不确定性人工智能的研究热点。

1. 混沌

在科学上，如果一个系统的演变过程对初态非常敏感，人们就称它为混沌（Chaos）系统[5]。混沌是决定性动力学系统中出现的一种貌似随机的运动，其本质是系统的长期行为对初始条件的敏感性，如我们常说的“差之毫厘，失之千里”。

1972 年 12 月 29 日，美国麻省理工学院教授、混沌学开创人 Edward N. Lorentz 在美国科学发展学会第 139 次会议上发表了题为《蝴蝶效应》的论文，提出一个貌似荒谬的论断：在巴西一只蝴蝶翅膀的拍打能在美国德克萨斯州产生一个龙卷风，并由此提出了天气的不可准确预报性。这位气象学家制作了一个电脑程序模拟气候的变化，并用图像来表示。最后他发现，图像是混沌的，而且十分像一只张开双翅的蝴蝶，因而他形象地将这一图形以“蝴蝶扇动翅膀”的方式进行阐释。时至今日，这一论断仍为人津津乐道，更重要的是，它激发了人们对混沌学的浓厚兴趣。今天，伴随计算机等技术的飞速进步，混沌学已发展成为一门影响深远、发展迅速的前沿科学。

一般地，如果一个接近实际而没有内在随机性的模型仍然具有貌似随机的行为，

就可以称这个真实物理系统是混沌的。一个随时间确定性变化或具有微弱随机性的变化系统，称为动力系统[6]，它的状态可由一个或几个变量数值确定。而一些动力系统中，两个几乎完全一致的状态经过充分长时间后会变得毫无一致，恰如从长序列中随机选取的两个状态那样，这种系统被认为敏感地依赖于初始条件。而对初始条件的敏感的依赖性也可作为一个混沌的定义。

与我们通常研究的线性科学不同，混沌学研究的是一种非线性科学，而非线性科学研究似乎总是把人们对“正常”事物、“正常”现象的认识转向对“反常”事物、“反常”现象的探索。例如，孤波不是周期性振荡的规则传播；“多媒体”技术对信息存储、压缩、传播、转换和控制过程中遇到大量的“非常规”现象产生所采用的“非常规”的新方法；混沌打破了确定性方程由初始条件严格确定系统未来运动的“常规”，出现所谓各种“奇异吸引子”现象等[7]。混沌的发现是对确定论最大的冲击。过去科学家们认为确定的系统只能产生确定的结果，绝对不会产生随机性，而把产生随机现象的原因归结为外部的影响。混沌科学却告诉我们，确定性的系统也可以产生随机的结果。非线性动力学中的“混沌”概念有着严格的科学定义，而人工智能以及其他一般科学中的混沌，指的是由确定性系统产生的复杂的随机行为，是人类社会和自然界中普遍存在的一种不确定运动形态。

混沌理论研究如何从貌似无序而实际有序、表面上看来是杂乱无章的现象中找出其确定的规律。生物学家发现在人类的心脏中有混沌现象存在，显微镜下交叉缠绕的微细血管，高度复杂中也有惊人的有序性；在生物脑神经系统中，从微观的神经膜电位到宏观的脑电波，都可以观察到混沌的形态，这说明混沌是生物神经系统的特性之一。

科学家给混沌下的定义是：混沌是指发生在确定性系统中的貌似随机的不规则运动，一个确定性理论描述的系统，其行为却表现为不确定性（不可重复、不可预测），这就是混沌现象。进一步研究表明，混沌是非线性动力系统的固有特性，是非线性系统普遍存在的现象。牛顿确定性理论能够完美处理的多为线性系统，而线性系统大多是由非线性系统简化来的。因此，在现实生活和实际工程技术问题中，混沌是无处不在的。从 Lorentz 第一次发现混沌现象至今，关于混沌的研究一直是科学家、社会学家、人文学家所关注的。研究混沌，其实就是发现无序中的有序，但今天的世界仍存在着太多的无法预测，混沌，这个话题也必将成为全人类性的问题。由于知识有限，我们在此只是做了较为简单的介绍和引入，希望有更多人能走进混沌之门，以更深邃的眼光来审视这个世界。

2. 分形

分形（Fractal）通常被定义为“一个粗糙或零碎的几何形状，可以分成若干个子部分，且每一部分都（至少近似地）是整体缩小后的形状”，即具有自相似的性质（见图 1.1）。

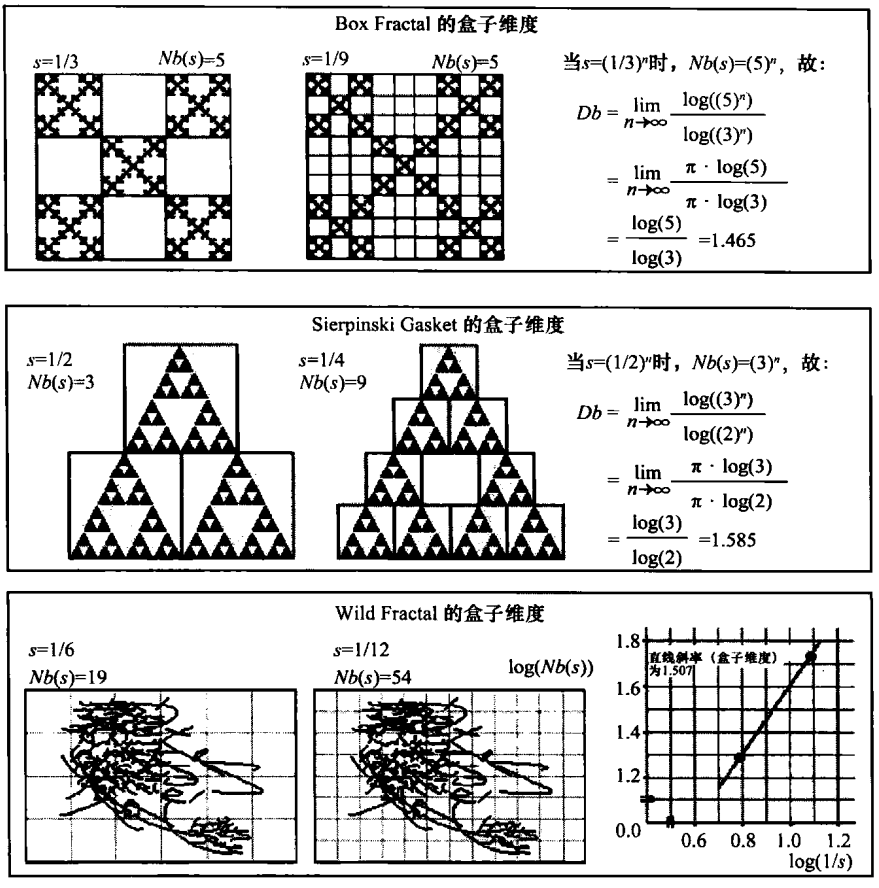


图 1.1 分形几何图

Fig.1.1 Fractal geometry

分形思想的根源可以追溯到 17 世纪，而对分形使用严格的数学处理则始于一个世纪后卡尔·魏尔施特拉斯、格奥尔格·康托尔和费利克斯·豪斯多夫对连续而不可微函数的研究。但是分形（Fractal）一词直到 1975 年才由本华·曼德博创造出，来自拉丁文 Frāctus，有“零碎”、“破裂”之意。

在 20 世纪 70 年代，法国数学家芒德勃罗（B. B. Mandelbrot）在他的著作中探讨了“英国的海岸线有多长”这个问题。这依赖于测量时所使用的尺度。如果用千米作测量单位，从几米到几十米的一些曲折会被忽略；改用米来做单位，测得的总长度会增加，但是一些厘米量级以下的就不能反映出来。由于涨潮落潮使海岸线的水陆分界线具有各种层次的不规则性。海岸线在大小两个方向都有自然的限制。取不列颠岛外缘上几个突出的点，用直线把它们连起来，得到海岸线长度的一种下界，使用比这更长的尺度是没有意义的。还有海沙石的最小尺度是原子和分子，使用更

小的尺度也是没有意义的。在这两个自然限度之间，存在着可以变化许多个数量级的“无标度”区，长度不是海岸线的定量特征，就要用分维。

数学家柯赫（Koch）也做过类似的试验。从一个正方形的“岛”出发，始终保持面积不变，把它的“海岸线”变成无限曲线，其长度也不断增加，并趋向于无穷大。可以看到，分维才是“Koch 岛”海岸线的确切特征量，即海岸线的分维均介于 1 到 2 之间。

这些自然现象，特别是物理现象和分形有着密切的关系，银河系中的若断若续的星体分布，就具有分维的吸引子。多孔介质中的流体运动和它产生的渗流模型，都是分形的研究对象。分形是为了表征复杂图形和复杂过程而引入自然科学领域的，它使自然界中普遍存在的螺旋、树状、斑纹、云彩、火焰等复杂不确定现象的研究变得简单，同时，在此基础上形成了分形几何学研究领域。

由于计算技术和计算机图形学的发展，使得大量的自然景物可以模拟。自然界、人类社会中存在大量的复杂不规则现象和无尺度的不确定现象，如湍流、复杂网络上的传播行为、金融市场的价格波动等。这类无尺度、自相似中的不确定性都是分形的研究对象。金融市场的价格变化，表面随机、无序，但是通过对大量现实数据的分析，科学家们发现价格变化存在无尺度、自相似的特点，用分形可以模拟价格随时间的变化，多重分形还可以描述市场的不确定性。与常规的统计方法不同，分形将复杂体系分解，可以体现复杂体系中的内部精细结构和蕴含的信息，而统计方法只能得到宏观的、粗略的估计。

分形理论既是非线性科学的前沿和重要分支，又是一门新兴的信息学科。作为一种方法论和认识论，其启示是多方面的：一是分形整体与局部形态的相似，启发人们通过认识部分来认识整体，从有限中认识无限；二是分形揭示了介于整体与部分、有序与无序、复杂与简单之间的新形态、新秩序；三是分形从一个特定的层面揭示了世界普遍联系和统一的深层次内涵。

3. 复杂网络

在网络理论的研究中，复杂网络（Complex Network）是由数量巨大的节点和节点之间错综复杂的关系共同构成的网络结构。用数学的语言来说，就是一个有着足够复杂的拓扑结构特征的图。复杂网络具有简单网络，如晶格网络、随机图等结构所不具备的特性，而这些特性往往出现在真实世界的网络结构中。复杂网络的研究是现今科学研究中的一个热点，与现实中各类高复杂性系统，例如互联网网络、神经网络和社会网络的研究有密切关系。

我国学者钱学森给出了复杂网络的一个较严格的定义：具有自组织、自相似、吸引子、小世界、无标度中部分或全部性质的网络称为复杂网络。复杂网络简而言之即呈现高度复杂性的网络。其复杂性主要表现在以下几个方面。

（1）结构复杂。表现在节点数目巨大，网络结构呈现多种不同特征。

（2）网络进化。表现在节点或连接的产生与消失，例如 World-Wide Network，网页或链接随时可能出现或断开，导致网络结构不断发生变化。

（3）连接多样性。节点之间的连接权重存在差异，且有可能存在方向性。

（4）动力学复杂性。节点集可能属于非线性动力学系统，例如节点状态随时间发生复杂变化。

（5）节点多样性。复杂网络中的节点可以代表任何事物，例如，人际关系构成的复杂网络节点代表单独个体，万维网组成的复杂网络节点可以表示不同网页。

（6）多重复杂性融合。即以上多重复杂性相互影响，导致更为难以预料的结果。例如，设计一个电力供应网络需要考虑此网络的进化过程，其进化过程决定网络的拓扑结构。当两个节点之间频繁进行能量传输时，它们之间的连接权重会随之增加，通过不断的学习与记忆逐步改善网络性能。

近年来，学界关于复杂网络的研究方兴未艾。特别是，国际上有两项开创性工作掀起了一股不小的研究复杂网络的热潮。一是 1998 年 Watts 和 Strogatz 在《Nature》杂志上发表文章，引入了小世界（Small-World）网络模型[8]，以描述从完全规则网络到完全随机网络的转变。小世界网络既具有与规则网络类似的聚类特性，又具有与随机网络类似的较小的平均路径长度。二是 1999 年 Barabási 和 Albert 在《Science》上发表文章指出，许多实际的复杂网络的连接度分布具有幂律形式。由于幂律分布没有明显的特征长度，该类网络又被称为无标度（Scale-Free）网络。而后科学家们又研究了各种复杂网络的各种特性。国内学界也已经注意到了这种趋势，并且也开始展开研究。

具有小世界效应和无尺度特性的复杂网络近年来备受人们关注。大量的真实网络，如互联网、万维网、电力网、航空网、食物链、人际关系网都是这种复杂网络。复杂网络研究的学者主要来自图论、统计物理学、计算机网络研究、生态学、社会学以及经济学等领域，研究所涉及的网络主要有：生命科学领域的各种网络（如细胞网络、蛋白质-蛋白质作用网络、蛋白质折叠网络、神经网络、生态网络）、Internet/WWW 网络、社会网络（包括流行性疾病的传播网络、科学家合作网络、人类性关系网络、语言学网络等）。他们使用的主要方法是数学上的图论、物理学中的统计物理学方法和社会网络分析方法[9-11]。

复杂网络拓扑结构的不确定性是复杂网络研究的基本问题。近年来研究发现，很多实际的复杂网络既不完全规则也不完全随机，而是介于完全规则和完全随机这两个极端之间，既具有类似规则网络的较大集聚系数，又具有类似于随机网络的较小平均路径长度，这就是小世界网络[11]。

混沌、分形和复杂网络都是研究不确定性的非线性科学，它们试图找出介于有序与无序、宏观与微观、整体与部分之间的新秩序。混沌、分形和复杂网络的不断研究发展为研究人类复杂的不确定行为提供了新的理论。

基于认识论的不确定性人工智能主要研究人类认知过程中的不确定性，主要包含人类感知的不确定性、记忆的不确定性与思维的不确定性。

感知的不确定性主要指人类在感知客观世界时，可以从毫不相关的粒度或者尺度上观察和分析同一问题，而且还能很快地从一个粒度世界跳到另一个粒度世界。感知的不确定性极大的体现了人类智慧的光辉魅力。

记忆的不确定性主要体现在人类思维的记忆信息反映。通过记忆过程，人脑存储了大量的知识与信息，为我们的心理发展、个性形成、社会活动提供必要的认知基础。但是记忆常常是不准确的，随着时间的推移，记忆中事物的面貌、过程的细节会变得模糊。但无论短时记忆还是长时记忆，大脑都会记取事物的整体和特征，忽略无关紧要的细枝末节。这样，人们会得到不具体的、模模糊糊的印象，但却能长时间地保持记忆，并且在需要的时候对记忆中的有用信息进行提取。

人类的思维有精确的一面，更有不确定的一面，尤其是涉及联想、创造、顿悟等形象思维时，更没有确切的规律可言。人类习惯于用自然语言进行思维，思维的结果往往是可能如何、大概如何等定性的结论。人类还擅长于通过幻想的、直觉的、抽象的、创造的形象思维来认识客观世界，几乎不可能像计算机一样做精确的运算或者严密的逻辑推理，和计算机相比，人脑并不具备快速、可靠的计算能力，或者海量数据的存储能力，但是这并不妨碍人们的学习和创造能力，不妨碍人们具有发达的、灵活的高级智能。可以说，在人类思维活动中，不确定的形象思维占据了绝大部分，与确定的、精确的符号思维活动相比，后者可以说是微不足道的。人类思维活动的这种定性特征，往往比定量计算更准确、更到位、更贴切。

自然语言的不确定性也是基于人类认知过程的不确定性人工智能研究的重要内容。语言，尤其是文字语言，是人类与其他一切生物在智能上的最大区别，也是人类智能最突出的表现。人类用语言描述和记载客观世界，描述和记载情感、心理和认知活动，因此，无论研究人类智能还是人工智能，应该从研究自然语言开始；研究不确定性人类智能和不确定性人工智能，也应该从研究自然语言的不确定性开始。

1.3.3 不确定性人工智能的主要研究方法

传统的人工智能以确定性作为基础，出现了以符号主义、连接主义、行为主义为代表的主流学派，在研究初期分别取得了一系列的理论和应用成果。随着研究的深入和相关学科的发展，人们逐渐意识到要想彻底解决人工智能研究中遇到的难题，需要考虑到不确定性的因素，把确定性和不确定性有机统一，制造出真正“智能”的机器。不确定性人工智能以不确定性为研究的重点和主要方面，从崭新的角度对人工智能进行研究，弥补了传统确定性人工智能的不足。

随机性和模糊性是不确定性的最基本内涵。经过多年的不断研究，众多学者先后提出了概率论、模糊集、粗糙集、云模型、分形网络、混沌等针对不确定性问题