1. 科学知识扩散的能力论（唐诗琪）

科学知识扩散能力是衡量科学知识创造、传播、流动和共享等程度的重要手段，是体现科学知识体系演化和变革水平的重要指标。作为科学知识扩散研究不可或缺的一部分，科学知识扩散能力汇聚着科学知识体系发展的内在动力。它通常作为科学知识扩散研究的因变量，描述科学知识扩散的统计特征，为科学知识扩散的理论研究提供定量方法和结果支持。科学知识扩散能力可以通过不同的维度进行测量，因此，科学知识扩散能力的不同研究方法和测度指标共同构成了科学知识扩散能力测度的方法体系。根据不同的测量需求，可以构建不同的测度指标，选择不同的测度方法，进而使用或开发典型的算法和分析工具对科学知识扩散能力进行度量。

1.1. 科学知识扩散能力概述

科学知识扩散是科学知识传播、发生动态演进的过程与现象，同时它也具备改造知识体系乃至科学系统的能力。由于受知识载体、传播环境等不同因素的复杂影响，科学知识扩散的能力存在较大差异。目前科学知识扩散的研究视角主要集中在学科之间的扩散、地理区域间的扩散以及知识随时间的扩散，随着对专利研究的深入，科学知识向技术的扩散也开始逐渐进入人们的研究视野。科学知识扩散这一主题已经成为多个学科的研究焦点。

1.1.1. 科学知识扩散能力释义

“能力”是指生命物体对自然探索、认知、改造水平的度量。为了表征科学知识扩散的变化程度和水平，科学知识扩散能力是指在某一特定条件下，推动科学知识扩散的各种环境因素和物质条件的总和。科学知识扩散的本质是科学知识单元的游离与重组，是从初级知识形态向高级知识形态演化，形成一定的知识体系和学科范式，并可能发展为科学变革的过程。广义的科学知识扩散包括知识单元在时间和空间双重维度的变化，包括内容层面的语义距离变化、结构层面的语义结构变化、空间层面的路径变化及方向变化、时间层面的时序变化等。狭义的科学知识扩散是知识单元从知识主体到知识客体的扩散过程，包括新知识的产生。总的来说，科学知识扩散是科学知识传播、发生动态演进的过程与现象，反映了科学知识体系自身的发展演化。科学知识扩散具有复杂的驱动力，在不同因素的交织影响下，科学知识扩散具有一定的自发性和自觉性，具备改造知识体系、学科范式乃至科学系统的水平。

科学知识扩散所造成的变化可以从不同的视角进行度量，同时受载体、传播环境等不同因素影响，科学知识扩散的快慢和扩散的范围等也会有所不同，即科学知识扩散能力的不同。在科学知识扩散能力定量测度指标体系中，速度、广度

和强度是常用的测度指标维度，它们的定义与科学知识扩散能力测度的研究视角和使用的测度方法直接相关。

1.1.2. 科学知识扩散能力分析视角

科学知识扩散可以通过不同的分析视角进行研究。常见的科学知识扩散分析视角包括空间扩散如学科（领域）之间的扩散、地理区域间的扩散，通常使用科学知识扩散的广度指标进行测量。时间扩散是指科学知识的时序扩散，通常使用科学知识扩散的速度指标进行测量。

科学知识在空间维度的扩散主要包括学科（领域）之间、科学与技术之间、区域物理距离等方面。其中，学科间的知识扩散是开展跨学科研究的重要前提和基础，探索不同学科之间的知识扩散过程、方式和机制等对跨学科的形成和发展规律等探究具有重要的意义。其中跨学科的“学科”可以通过科学文献所属学科来表征，就技术知识而言，则通过专利 IPC 号来表征知识所属的不同领域。

科学知识通过不同的形式在不同学科或者技术领域之间扩散。由于引文数据相对容易获取，不同学科之间科学文献的引证关系是研究科学知识在学科之间扩散的最常见方式，文献的引用往往意味着其他学科或者研究领域的理论或方法被嵌入到当前学科中；合作是科学知识在学科间扩散的另一种方式，即不同学科的研究人员通过合作发表科学文献和专利成果，该成果往往融合了不同学科领域间的知识；更为细化的研究则探究不同知识单元在引用过程中性质的变化，如某一数学原理是否在其他学科领域作为方法被加以应用，从而探索不同学科的知识需求以及“跨学科”的科学所以形成的知识基础。其他形式的扩散还包括人员流动等，但以研究人员为载体进行科学知识在科学间扩散的研究需要明确相关人员的学科或教育背景，因而往往受到数据获取方面的限制。

在知识经济时代，科学知识扩散成为影响区域经济竞争力强弱的关键因素之一。认识区域间的科学知识扩散机制有利于提高区域创新系统的运行效率，从而促进区域经济的健康快速发展。不同于一般意义上的科学知识传播，区域科学知识扩散往往属于社会经济活动的伴随行为，与不同主题的社会经济交往密切相关。区域科学知识在地理区域之间的扩散可以存在于个体或团体之间，如科学文献或专利等的合作发表或彼此引用；也可以存在于组织机构间，如大学或研究机构与公司企业之间的合作；相对隐性的机构间知识扩散也包括了人员的流动。相对于科学文献，以专利为载体的科学知识扩散具有更加多样的渠道，诸如专利转让、专利许可等商业行为促进了企业的知识吸收与知识重组，也是知识在不同区域机构或企业间扩散的重要形式。以上共同构成了科学知识的地理区域扩散。

伴随着知识不断被创造，知识内容和学科领域研究主题等也在不断发生继承和演化，即科学知识的时间扩散。对科学知识时间扩散的早期研究往往依赖于专家总结，近年来随着技术的发展则能够通过大数据与人工智能手段对学术论文和

专利等的研究主题的发展脉络和未来趋势进行分析预判。

以科学文献和专利为主体的引文网络挖掘和时序分析可以有效揭示知识的时间扩散与演进过程。所谓时序网络，即将各个节点按其产生时间排列，由此可以分析得出知识扩散过程中的基础文献和关键文献等。CiteSpace 软件中的时间线图等功能与此类似。对科学知识扩散的更细粒度分析被称为主题演化，以知识单元作为知识的载体。知识单元的识别可以使用论文关键词或作者关键词直接进行表征，也可以通过 LDA 主题模型等从文献标题、摘要等文本中提取。在主题的时序关联方面，除通过构建引文网络进而提取引文主题主路径进行分析外，通过主题文本相似度等进行主题的关联测度也开始逐渐成为新的研究趋势。

科学与技术之间的知识扩散一直以来都是科技管理、科学计量学相关领域研究的热点问题。已有研究如科学-技术关联模型已经表明了科学与技术之间相对独立又相互关联的复杂关系，科学知识是推动技术发展的前提和基础，技术需求又催生科学知识的进一步发展，二者共同促进科学发展和技术创新。因而探究科学知识如何促进技术创新以及科学与技术的关联关系等问题，对于科学发现规律的探索、未来科学前沿发展和研究主题演化预测等都具有重要的意义。

1.2.科学知识扩散能力的研究方法

科学知识扩散能力的研究方法包括了定性分析的研究方法和定量分析的研究方法。其中定性方法主要以直接分析法和德尔菲法为主，通过专家评议从主观上评价科学知识扩散的程度和水平。随着数据驱动和实证主义研究范式的发展，以及对科学知识扩散的交叉学科研究属性的认识，更多研究者开始融合自然科学和社会科学的研究视角，基于数学、物理、计算机与情报学等多学科方法的移植和创新实现对科学知识扩散能力的定量研究。典型的科学知识扩散能力测度方法包括基于不同网络及其特征的方法、基于指标特征的统计方法、机器学习与文本挖掘方法、系统动力学方法等。

1.2.1. 科学知识扩散能力研究的演进

早期科学知识扩散的研究主要以定性分析方法为主，如直接分析法和德尔菲法等，通过专家评议从主观上评价科学知识扩散程度。关于知识扩散的研究也往往在社会学、管理学等一系列相互独立的学科领域间展开。

随着数据驱动和实证主义研究范式的发展，科学知识扩散作为交叉学科的研究属性不断强化，对科学知识扩散能力的研究开始从单纯社会科学向融合自然科学与社会科学不同学科领域的方向转变。基于数学、物理、计算机、情报学与管理学等多学科方法的移植和创新对科学知识扩散能力的定量研究方法得到快速发展。

在定量测度方面，统计学方法成为科学知识扩散能力研究主要手段。如应用相关性和回归等统计学方法可以探寻不同影响因素对科学知识扩散能力的正向或负向影响。在科学知识扩散能力的具体测度方面，已有研究建立了包括扩散广度、宽度、速度等方面在内的一系列科学知识扩散指标用于科学知识扩散能力的度量。

文本挖掘也是科学知识扩散能力测度的重要手段之一。随着大数据时代的到来，如今的研究者可以获得比纸质书籍和期刊研究者更多的数据信息，诸如各种电子数据库中保存着的期刊全文文本数据、出版数据、引用数据乃至其他研究者的评价数据等等。如何从广阔的文本数据中挖掘出所需信息用于科学知识扩散能力的研究，新兴的各种机器学习和深度学习理论模型为之提供了解决方法。

此外，复杂网络分析在科学知识扩散能力研究中也有重要应用。科学知识或者其载体本身是一项个体，知识与知识之间又具有多样的联系。这一特性使得科学知识扩散自然依托于知识所构建的复杂网络系统。与之相关的，如社会网络理论等，提供了从“关系”和“空间”研究科学知识扩散能力的重要视角；系统动力学则从科学知识扩散系统内部的机制、微观结构等入手，对系统进行剖析与建模，借助计算机仿真模拟技术，分析科学知识扩散的内部结构与其动态行为之间的关系。

1.2.2. 科学知识扩散能力研究的体系

科学知识扩散研究具有典型的交叉学科研究属性，科学知识扩散能力的研究包含定性和定量两方面的研究方法。对科学知识扩散能力的构成要素、影响机理、作用路径等基础理论问题多以定性分析方法为主。在对科学知识扩散能力的测度方面主要以定量研究手段为主。特别是在科学知识爆炸式发展以及社交媒体快速兴起的科学知识交流新范式下，大数据驱动的实证主义研究范式已经成为知识扩散能力测度研究的主要途径。相应的研究方法体系、定量测度指标体系以及典型算法与分析工具等都得到快递发展。

实证主义的研究范式注重研究客观事实和社会产物，将客观存在的社会现象作为研究起点，重视对社会规律进行科学概括，试图寻求社会现象间的相关关系或因果关系。它关注被研究对象的一般性、普遍性或规律性。实证研究方法包括质化研究，如扎根理论、现象学研究、案例分析等；在科学知识扩散能力研究中更多被使用的则是量化研究，主要通过随机抽样调查方法去搜集资料，包括问卷法、结构性观察法、问卷访谈法等；倾向于运用诸如统计图表类的定量技术或利用统计软件和计算机去处理、分析资料，以及用公式、数量模型去表达经得起检验的假设；既使用了包括观察、实验、测量、演绎、假说等自然科学的或经验科学的研究方式，还使用了包括逻辑的、数学的、统计的分析方法。此外，质化和量化的混合研究超越传统质化和量化方法之间关于现象归纳演绎、因素主观客观

之间的争论，是日益受到重视的综合研究方法。

当前大数据驱动的实证主义研究范式多用于横向研究，常常围绕某种社会现象（事件）、社会问题而不是针对某一个时期内去搜索资料，适用于对大范围的社会活动结果或大量的社会现象（问题）的发生作宏观及微观研究分析。常运用统计学原理与方法，把大量社会现象的产生及演变，视为一种随机现象进行研究，具体归结为对随机事件和随机变量的演变趋势和规律的研究，其基本元素包括理论分析、研究假设、样本数据、描述性统计以及多元回归等统计学分析等。其相较于案例研究的优点在于，在数据样本足够的情况下可以通过变量控制推断因素与结果之间的因果关系，排除其它因素的干扰。以此方法对社会现象开展研究，则存在一些常见问题需要使用者在进一步分析中加以考虑。内生性问题，主要由测量误差、遗漏解释变量、因素互为因果三个方面的原因造成，与之相对的方法包括工具变量法、倾向得分匹配、双重差分等；更广泛的稳健性检验问题，如模型运用方法替换、排除其它理论假说与多重共线性问题的检验等。视研究研究问题的不同，也可以开展分样本检验、调节效应分析、中介机制探索等，用以支持研究者的观点和假说，并进一步对所探索的社会问题开展讨论和思考。

1.2.3. 科学知识扩散能力的测度方法

科学知识扩散能力的测度方法主要有基于不同网络及其特征的方法、数理统计方法、文本挖掘相关方法以及系统动力学方法。其中，基于不同网络及其特征的方法是科学知识扩散能力的典型测度方法之一。根据扩散载体和扩散渠道的不同可对知识扩散网络进行划分。在根据知识载体作为网络节点和知识扩散渠道作为节点关系构建网络的基础上进行网络特征测度，是目前科学知识扩散测度的代表性研究范式之一，如通过计算相关网络的规模、密度、中心化程度等可以获得对学科领域、区域创新能力等的整体认识；通过观察网络内部结构随时间的变化，可以进一步分析知识扩散未来的发展趋势。

对于网络，常用的分析方法包括：图论、复杂网络分析、社会网络分析等，主要涉及的网络特性及其定义[1]如下表所示。

表 1 网络特性及定义

网络特性	定义	其它
网络密度（Density）	一个包含N个节点的网络的密度定义为网络中实际存在的边数M与最大可能的边	$\rho = \frac{M}{\frac{1}{2}N(N-1)}$

[1] 汪小帆等, 网络科学导论. 2012: 高等教育出版社.

	数之比	
特征路径长度（characteristic path length）	网络的特征路径长度定义为有N个节点的网络中任意两个节点i和j之间的距离d_{ij}（两个节点之间的距离为连接这两个节点的最短路径上边的数目）的平均值	$$L = \frac{1}{\frac{1}{2}N(N-1)}\sum_{i \geq j} d_{ij}$$
连通性（connectivity）	网络的每一对顶点之间至少都存在一条路径，则称网络连通	
聚类系数（clustering coefficient）	网络中一个度为k_i（即有k_i个有边直接相连的邻居节点）的节点i的聚类系数C_i定义为节点i的k_i个邻节点之间实际存在的边数E_i与节点的k_i个邻接点两两互为邻居情况下边数之比	$$C_i = \frac{E_I}{\frac{1}{2}k_i(k_i-1)}$$
度分布（degree distribution）	度分布P(k)定义为网络中一个随机选择的节点的度为k的概率	常见的度分布如正态分布、长尾分布、幂律分布等
中心性（centrality）	用于刻画节点在网络中所处位置	包括度中心性（degree centrality）、介数中心性（betweeness centrality）、接近中心性（closeness centrality）和特征向量中心性（eigenvector centrality）

数理统计是科学知识扩散能力测度的基础方法，其以概率论为基础，研究所收集数据样本的相关特征，并由样本特征推断总体特征。具体方法包括描述性统计方法，如通过平均数、中位数、分位数等描述数据的集中趋势和平均水平，通过方差、标准差、变异系数等反应数据的差异程度，通过二项分布、正态分布等常见概率分布对知识扩散能力情况进行简单刻画；推断性统计方法主要包含参数估计与假设检验两部分：根据已有样本数据，给出总体分布的未知参数或是根据样本观测值对给定两种相斥假设做出可信度判断；相关分析与回归分析方法则更多被用于探究科学知识扩散能力及其影响因素之间的关系，常见的有皮尔森相关

系数、斯皮尔曼相关系数、线性回归等；聚类也是一种常用方法，常被用于知识的分组和分布状况的度量。

通过文本挖掘量化知识扩散特征是一种新兴的科学知识扩散能力测度方法。虽然该方法的大规模应用得益于计算机的普及和机器学习等概念的兴起，但其具有悠久的数学历史，相关语言模型的统计学基础可以追溯到对词和词项频率的计数如词袋模型、利用齐普夫定律预测词的出现概率以及通过 TF-IDF 向量表示词的重要程度等。一项大的发展在于从自然语言中提取结构化的数值数据即向量，这使得更多的数学工具和方法可以被使用，计算机也因此能够解释和存储语句的“含义”，自然语言处理进入语义时代。向量化的方法从 one-hot 编码、潜在语义分析（Latent Semantic Analysis，LSA）、word2vec 到现在研究人员常使用的 BERT 等不一而足，结合深度学习中的卷积神经网络（CNN）、循环神经网络（Recurrent Neural Network,RNN）、长短时记忆神经网络（Long Short-Term Memory, LSTM）、序列到序列建模（seq2seq）以及注意力机制（attention）等，可以实现多样的任务如序列标注、文本分类。在科学知识扩散能力测度研究中常出现的主题词共现网络，其构建所需的主题通常使用隐狄利克雷分布（Latent Dirichlet Allocation, LDA）或者长短时记忆神经网络结合条件随机场（Long Short-Term Memory-Conditional Random Field, LSTM-CRF）的方式从海量文本中加以提取。

系统动力学方法也是科学知识扩散能力的测度方法之一，其优势在于通过模拟仿真技术，研究者可以对科学知识扩散这一复杂系统的演变和相关影响因素进行分析。相关模型如库存(stock)和物流图表展现了系统内部运行方式，因果环图(casual loop)展示因果关系和结构中主要的反馈环路等；通过仿真长时间“运行”系统，可以模拟真实情况下系统的发展演变，预测科学知识的扩散情况；通过灵敏度分析可以展现系统的关键的平衡点和最优条件，有助于研究人员分析和理解科学知识扩散能力达到最大时的系统情况，对实际应用具有一定的指导意义。

1.3.科学知识扩散能力的测度指标

在科学知识扩散能力的研究中，对扩散能力的测度是当前学界和实践领域最为关注的问题之一。通过构建不同的测度指标以及指标体系对科学知识扩散能力的不同维度进行测量和评价，其中速度、广度和强度是科学知识扩散能力定量测度指标体系中最常用的指标，它们的定义与科学知识扩散能力测度的研究维度和使用的测度方法直接相关。一般而言，广度反映科学知识扩散的范围，强度反映科学知识扩散的深度，速度反映科学知识扩散的快慢程度。除此之外，其他的测度指标还包括延时扩散、睡美人以及替代计量学新范式下的传播热度等。

1.3.1. 科学知识扩散测度指标体系的构建

指标是衡量目标的单位。指标体系构建是指将多个指标按照一定关系构建成一个整体。一般而言，指标包括独立指标和复合指标，其中复合指标可以通过公式拆解成独立指标。构建科学知识扩散测度指标体系能够将科学知识扩散逻辑化，有助于对科学知识扩散能力的评价。

1. 测度指标体系的构建原则

在科学知识扩散测度指标体系构建过程中遵循指标体系拆解的原则，如MECE 原则（相互独立，完全穷尽）、CSCE 原则（完备性、系统性、可执行性、可解释性），目的是保证指标的价值型以及指标之间的相互独立性。同时就单独指标的构建而言，遵循系统性原则，典型性原则，动态性原则，简明科学性原则，可比、可操作、可量化原则以及综合性原则等，充分考虑指标在时间维度上可能具有的动态变化、数据是否易获取以及计算过程是否简明易懂，进而有助于对科学知识扩散的综合分析与评价。

2. 测度指标体系的构建过程与方法

根据表现维度对科学知识扩散指标体系进行横向拆解，将其转化为科学知识扩散广度、强度与速度。在此基础上根据科学知识扩散的不同载体与不同表现进行纵向的细分拆解，针对学科、期刊、文献、知识单元等分别构建指标，进而建立层次指标体系。在纵向拆解建立指标的过程中，则通过可获得的数据作为统计资料的支持、对典型案例进行分析以构建具体指标。

1.3.2. 科学知识扩散广度指标

广度指的是事物的范围，用以表征丰富程度和多元化程度。科学知识扩散广度是从覆盖范围的角度对扩散进行描述，科学知识扩散的覆盖范围越大，意味着扩散的对象更加丰富和多元，扩散的广度越大。由于知识扩散是一个持续的过程，因此知识扩散广度指标应当能够反映一定时间内知识扩散的累积状况。

表 2 科学知识扩散广度指标

指标	载体	提出者/修正者	年份	定义
JDF	期刊	Rowlands[2]	2002	$\mathrm{JDF}=\frac{\mathrm{U}*100}{\mathrm{Cit}}$ 其中 U 为被引涉及的不同期刊的数量，Cit 为被引次数。在指定时间范围内，对某期刊的每 100 次被引涉及的不同期刊的数量即为该期刊的扩散因子。

[2] Rowlands and Ian, Journal diffusion factors: a new approach to measuring research influence. Aslib Proceedings, 2002. 54(2): p. 77-84.

NewJDF	期刊	Frandsen[3]	2004	$JDF = \frac{U}{Pub}$ 其中 U 为该期刊被引涉及的不同期刊的数量，Pub 为载文量。指定时间范围内某期刊刊载的论文中平均每篇涉及的不同施引期刊的数量即为该期刊的扩散因子。
FDB	文献	Liu[4]	2010	对于指定的一组论文，施引论文所属的 ESI 学科数量就是学科扩散广度。
跨学科引用指数	学科	Porter, A L Chubin, D E[5]	1985	某学科引用（或被引）的所有文献中引用（或被引）其他学科文献的比例。

期刊扩散指标（Journal Diffusion Factor, JDF）最早受到 JIF 的启发，用于科学知识扩散广度的测度。但由于一个学科范围内期刊的数量有限，因此采用 JDF 衡量期刊扩散指数时期刊的被引次数会造成较大的影响。因此产生了其改进 NewJDF，对期刊载文量进行了控制。除期刊外该指标也可用于单篇文献、作者、学科等知识扩散载体。学科扩散广度（Field Diffusion Breadth, FDB）基于科学论文的学科分类定义，该指标也可应用于专利，扩散广度即为施引专利所包含的类别，通过 IPC 类别进行表征。

总的来说，扩散广度一般被定义为某一知识载体在其他不同层次载体中的被引次数。同时为了平衡载体数量本身的不均衡性，如不同学科包含的期刊数量不同、不同期刊的载文量不同、不同文献内部知识单元数量不同等，可以通过比值等归一化形式对广度指标进行控制。

1.3.3. 科学知识扩散强度指标

科学知识扩散强度是从频次的角度对扩散进行描述，知识的扩散次数越多，科学知识扩散的强度越大。相较于扩散广度，强度更加强调扩散源自身向外扩散知识的情况。一般而言，科学知识扩散强度所指的是指定路径上的扩散频次。但知识扩散的路径很可能并非单一线性，而是多级发散的树状形式。考虑到多层级扩散的情况，深度作为科学知识扩散强度的一种形式也应纳入表征。

表 3 科学知识扩散强度指标

指标	载体	提出者/修正者	年份	定义
DINT	文献	Liu 和 Rousseau[6]	2010	对于指定的一组论文，某个 ESI 学科范围内的施引论文的数量就是这组论文在该

[3] Frandsen, T.F., Journal Diffusion Factors – a measure of diffusion? Aslib Proceedings, 2004. 56(1): p. 5-11(7).
[4] Liu, Y. and R. Rousseau, Knowledge diffusion through publications and citations: A case study using ESI-fields as unit of diffusion. Journal of the American Society for Information Science and Technology, 2010. 61.
[5] .Porter, A.L. and D.E. Chubin, An indicator of cross-disciplinary research. Scientometrics, 1985. 8(3): p. 161-176.
[6] Liu, Y. and R. Rousseau, Knowledge diffusion through publications and citations: A case study using ESI-fields as

ESI 学科范围内的扩散强度。

总的来说，扩散强度一般被定义为某一知识载体在同一层次知识载体中的被引次数，例如对于单篇科学文献或者专利而言，研究者往往统计其被其他文献引用的情况，以施引文献的被引频次作为科学知识扩散的强度指标。同时扩散广度和扩散强度之间也存在一定的关联。一般而言，两者呈反比例关系。扩散强度与扩散广度的乘积代表了知识载体的总体被引次数。即知识扩散强度(Field Diffusion Intensity, DINT)与知识扩散广度（FDB）之间存在如下关系：

$$T = \sum_{j=1}^{FDB} (DINT)_j$$

其中 j 指学科数量，T 指总被引次数。

以上的扩散强度是基于单一来源知识载体进行考量，如果考虑知识扩散的级联效应，即知识在产生一次扩散后，其扩散结果再次扩散，则情况又有所不同。知识的扩散路径形成网状或树状的结构（一般为有向无环图），深度、宽度、路径距离等也成为描述知识扩散的重要指标。知识级联包括两个特征：结构感染性和结构流行度，分别代表知识在扩散结构上的感染能力和知识扩散在整体上的流行程度，它们为从微观结构上观察知识扩散提供了新的视角。

表 4 科学知识级联扩散强度指标

特征	指标	提出者/修正者	年份	定义
结构感染性	维纳指数	Goel Sharad Anderson Ashton[7]	2015	图中任意两个节点之间最短路径的平均长度。 $v(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$ 其中，d_{ij}是节点 i 与节点 j 之间的最短路径长度。
	最大传播的相对规模			树中任意单个节点所拥有子节点数量的最大值占整棵树总节点数量的比重。
	父节点概率			任意两个节点拥有不同父节点的概率。
	平均深度			树中节点深度的平均值。

unit of diffusion. Journal of the American Society for Information Science and Technology, 2010. 61.

[7] Goel, S., et al., The Structural Virality of Online Diffusion. Management Science, 2015. 62(1): p. 180-196.

				$\text{Average depth} = \frac{1}{\lvert T\rvert - 1}\sum_{v\in T, v\neq r} d(v,r)$ 其中，T 是由 r 触发的级联，d(v,r)是节点 v 和 r 之间的距离。
结构流行性	一代数量			源节点第一代子节点的数量。
	二代数量	Hu Xiaojun Rousseau Ronald[8]	2016	源节点第二代子节点的数量。
	级联规模	闵超 Ding, Ying 等[9]	2018	扩散树的节点总数。

深度是最传统的结构感染性指标，但其没有考虑到过长的单链式扩散对测度造成的影响。一种改进方式是使用平均深度指标，但如果源节点具有很多一代节点，而一代节点则以较长的单链形式扩散，它的平均深度指标将会较高，但其影响力或许不如代际较少，但多点式扩散的节点。源自数学化学的指标维纳指数对此继续改进，考虑任意节点之间的最短距离，将整个扩散过程分化成为小的分支事件以完成对结构感染性的连续性测度。其它指标还包括最大传播的相对规模以及节点拥有不同父节点的概率等。

在结构流行性中，第一代子节点的数量直接表征了源知识的影响力。其相当于目前测度科学影响力的常用指标——直接引文次数。然而除直接影响外，知识也具有间接影响，如部分研究包含重要的基础性科学发现，其价值并未反映在自身的一代引文上，但造就了一些有影响力的跟进研究。因此二代数量也被纳入考量。级联规模考虑了不同代际节点的总影响以刻画知识扩散的结构流行性，相较前两者更好地反映了知识扩散的能力。需要注意的一点是级联规模并未真实反映源知识的影响力，因为其它代际的节点被赋予了一代节点一样的权重，而二者的重要性并不相等。为了更准确的度量，可以按照代际距离的远近为不同节点赋予不同的权重，能够更好地测度科学知识的扩散。

[8] Hu, X. and R. Rousseau, Scientific influence is not always visible: The phenomenon of under-cited influential publications. Journal of Informetrics, 2016. 10(4): p. 1079-1091.

[9] 闵超等, 单篇论著的引文扩散. 情报学报, 2018. 037(004): 第 341-350 页.

1.3.4. 科学知识扩散速度指标

速度描述的是单位时间内运动距离。科学知识扩散速度是从单位时间运动距离的角度对扩散进行描述，对于指定的扩散目标，所耗时间越短，则科学知识扩散速度越快。知识扩散速度会影响知识进步和知识创新，更快的知识扩散速度代表更快捷的知识共享与融合，进而加快科学知识的发展。

表 5 科学知识扩散速度指标

指标	载体	提出者/修正者	年份	定义
ADS	文献	Rousseau[10]	2010	$\mathrm{ADS}=\frac{\mathrm{FDB}}{\mathrm{Y_{pub}}}$ 其中，$\mathrm{Y_{pub}}$指文献年龄，FDB 指引用该论文的期刊或者 ESI 学科数量。即引用论文的 ESI 学科数量与论文年龄的比值。
引文滞后	文献	Nakamura[11]	2011	施引论文的发表时间减去被引论文发表时间所得的时间差。
半衰期	期刊			期刊在指定时间段内所有论文被引用的平均引文滞后的中值。
引文速度	文献	Wang, J Thijs, B 等[12]	2014	$\text{Citation speed}=\frac{\sum_1^{n-1}\mathrm{C_i}/\mathrm{C_n}}{n-1}$ 其中$\mathrm{C_i}$是文献发表后第 i 年的累计引文数量，n 是文献发表至统计时的时长（单位：年）。
引文延时	文献	Jian, W[13]	2013	$\text{Citation delay}=1-\text{Citation speed}=1-\frac{\sum_1^{n-1}\mathrm{C_i}/\mathrm{C_n}}{n-1}$

平均扩散速度（Average Diffusion Speed, ADS）描述了文献在其上一层载体，如期刊和学科中的扩散。接下来的四个指标则是单一的，与领域无关。与 ADS 相比，引文滞后揭示了微观的知识扩散速度，即知识从一篇文献扩散到另一篇文献的过程。引文速度和引文延时则着眼于整体，测度文献从发表到某个时刻整体上吸引引文的累积速度和延时。引文速度值越接近 1，累积速度越快，扩散越快；引文延时值越接近 1，引文延迟程度越高，扩散越慢。引文速度和引文延时的另一项重要作用是作为“睡美人”等具有特殊的延时扩散现象文献的判别指

[10] Liu, Y. and R. Rousseau, Knowledge diffusion through publications and citations: A case study using ESI-fields as unit of diffusion. Journal of the Association for Information Science & Technology, 2010. 61(2): p. 340-351.
[11] Nakamura, H., et al., Citation lag analysis in supply chain research. Scientometrics, 2011. 87(2): p. 221-232.
[12] Wang, J., B. Thijs and W. Glnzel, Interdisciplinarity and impact: Distinct effects of variety, balance and disparity. Working Papers of Department of Management, Strategy and Innovation, Leuven, 2014.
[13] Jian, W., Citation time window choice for research impact evaluation. Scientometrics, 2013. 94(3): p. 851-872.

标。

1.3.5. 科学知识扩散的其他指标

知识并非一经产生就能够得到传播，通常情况下普遍存在延时扩散的情况。就文献而言，从累积引文的视角出发，通过累积引文的曲线特征科学文献的生命周期可以被分为三个阶段：沉睡期、苏醒期和衰老期。侯剑华和张雪雯等从沉睡强度和苏醒强度两个维度构建累积引文睡美人指数（Cumulative citation sleeping beauty index,Cc index），也反映了延时扩散的科学知识扩散情况。沉睡强度越弱，苏醒能力越强，则科学知识扩散能力越强；沉睡强度越强，苏醒能力越弱，则科学知识扩散能力越弱。

表 6 睡美人相关指标

指标	载体	提出者/修正者	年份	定义	其它
沉睡强度	文献	侯剑华 张雪雯[14]	2020	沉睡强度衡量文献沉睡时间长短与沉睡期被引程度。文献沉睡程度 S 定义为：$$S=\frac{t_0}{y_0}$$其中，t_0表示文献苏醒时刻，y_0表示文献苏醒时刻累积引文量。	文献沉睡时间越长，累积引文越低，则文献的沉睡程度越强。
苏醒强度	文献	侯剑华 张雪雯	2020	苏醒强度衡量文献苏醒曲线的上升幅度。文献的苏醒强度 W 定义为：$$W=\frac{k-y_0}{t_2-t_0}$$即睡美人文献苏醒后，在苏醒期t_2-t_0内，文献累积引文量为$k-y_0$。	文献苏醒期越短、累积引文越高，则文献的苏醒强度越大。

[14] 侯剑华与张雪雯, 基于累积引文的科学睡美人识别方法研究. 情报学报, 2020. 39(9): 第 14 页.

知识在社交媒体中的扩散则与替代计量学相关。与传统的知识通过引文扩散不同，知识在社交媒体以不同的形式进行活动与交互。

表 7 社交媒体扩散相关指标

指标	含义	相关平台
阅读	阅读指标是基于网络上对学术成果的阅读行为产生的数据所构建的指标及其衍生指标。	出版商网站、期刊主页、学术社交网站如 ResearchGate、Mendeley 等。
下载	下载指标是基于网络上对学术成果的下载行为产生的数据所构建的指标及其衍生指标。	出版商（Springer、Elsevier、Wiley）、引文数据库（Scopus、WOS、CNKI）、期刊网站、专业网站、学者个人主页等。
收藏	收藏指标是基于网络上对学术成果的收藏行为产生的数据所构建的指标及其衍生指标。	社交媒体如推特、微信、学术平台如 Mendeley 等。
分享	分项指标是基于网络上对学术成果的分享行为产生的数据所构建的指标及其衍生指标。	分享来源可以是任意平台，分享指向平台一般是博客、facebook、微博等社交媒体平台。
提及	提及指标是基于网络上对学术成果的提及行为产生的数据所构建的指标及其衍生指标。	新闻报道平台、政策文件发布平台、微博、推特、学术会议等。
评论	评论指标是基于网络上对学术成果的评论行为产生的数据所构建的指标及其衍生指标。	YouTube、推特、同行评议、F1000、Publons 等。
复用	复用指标是基于网络上对学术成果的复用行为产生的数据所构建的指标及其衍生指标。	Github（分享代码、软件包）、Slideshare（分享幻灯片）等。

相较于基于引文的知识扩散，在社交媒体平台的知识扩散中，知识主体吸收知识的行为更加多样。其中阅读是最基础的知识吸收行为，经过阅读读者能对是否进一步利用学术成果做出判断；一般而言，在阅读后如果对知识感兴趣，则可能发生下载行为；收藏是读者对自身感兴趣知识的过滤并保留其被进一步利用的可能性；分享则是相对直接的扩散行为，表明发起分享者认为知识具有一定的价值；提及的应用场景则十分广泛，包括宣传、科普教育等；评论则是对知识较为

深入的讨论；复用则是对知识的认可和再利用[15]。基于以上指标可以分析科学知识在正式或非正式科学交流系统中的传播和扩散方向与程度。因而在替代计量学这一新范式研究背景下，科学知识扩散能力不仅可以通过引证进行度量，也可以通过其他知识吸收行为设计指标对知识扩散的广度、强度、速度乃至其他维度进行测度，从而衡量科学知识在社交媒体中基于不同吸收行为的扩散能力。

1.4.科学知识扩散能力的测度工具

受大数据驱动研究范式以及科学知识扩散的交叉学科研究属性的影响，科学知识扩散能力的测度算法与工具表现出多学科性、智能化、多样化等特征。其中典型算法主要包括网络相关算法、数理统计方法、机器学习、深度学习与自然语言处理相关算法、系统动力学相关算法等。主要的分析工具涵盖数据存储相关的数据库与查询语言、计算机辅助的计算框架、网络与计量相关的可视化、系统动力学相关的模拟仿真等诸多方面。

1.4.1. 扩散测度的典型算法

前文介绍了几种典型的科学知识扩散能力测度方法，不同方法涉及不同的属性与模型。算法是对这些属性的具体计算或模型实现。下面对科学知识扩散能力测度中较多涉及的一些典型算法进行介绍。

（1）网络相关算法

在基于网络的科学知识扩散研究中，度量节点的中心性是一项重要的研究问题，通常被用于知识扩散中关键文献或专利的识别。经典的中心性算法（或称排名算法）有 Page Rank 和 HITS 等。

Page Rank 一开始作为一种互联网网页重要度的方法被提出，后来被应用到社会网络分析等多个问题，其基本思想在于：如果网页 T 存在一个指向网页 A 的连接，则表明 T 的所有者认为 A 比较重要，从而把 T 的一部分重要性得分赋予 A。这个重要性得分值为：PR（T）/L(T，其中 PR（T）为 T 的 PageRank 值，L(T)为 T 的出链数。A 的 PageRank 值为一系列类似于 T 的页面重要性得分值的累加。即一个页面的得票数由所有链向它的页面的重要性来决定，到一个页面的超链接相当于对该页投一票。一个页面的 PageRank 是由所有链向它的页面（链入页面）的重要性经过递归算法得到的。一个有较多链入的页面会有较高的等级，相反如果一个页面没有任何链入页面，那么它没有等级。

在 HITS 算法（Hyperlink-Induced Topic Search）中，每个页面被赋予两个属性：hub 属性和 authority 属性。同时，网页被分为两种：hub 页面和

[15] 余厚强, 替代计量学. 2019: 科学技术文献出版社.

authority 页面。hub 页面指那些包含了很多指向 authority 页面的链接的网页，比如国内的一些门户网站；authority 页面则指那些包含有实质性内容的网页。HITS 算法的目的是：当用户查询时，返回给用户高质量的 authority 页面。其建立假设有：一个高质量的 authority 页面会被很多高质量的 hub 页面所指向；一个高质量的 hub 页面会指向很多高质量的 authority 页面。质量则由每个页面的 hub 值和 authority 值确定。其确定方法为：页面 hub 值等于所有它指向的页面的 authority 值之和;页面 authority 值等于所有指向它的页面的 hub 值之和。

路径发现算法通常被用于寻找科学知识扩散在网络中扩散的具体路径，进而可以对其扩散能力进行测度。深度优先和广度优先是两种经典的图路径发现算法。二者都是从起点开始顺着边搜索，此时并不知道图的整体结构，直到找到指定节点(即终点)。在此过程中每走到一个节点，就会判断一次它是否为终点。区别在于广度优先搜索会根据离起点的距离，按照从近到远的顺序对各节点进行搜索。而深度优先搜索会沿着一条路径不断往下搜索直到不能再继续为止，然后再折返，开始搜索下一条路径。在广度优先搜索中，有一个保存候补节点的队列，队列的性质就是先进先出，即先进入该队列的候补节点就先进行搜索。在深度优先搜索中，保存候补节点是栈，栈的性质就是先进后出，即最先进入该栈的候补节点就最后进行搜索。对于非加权图，广度优先搜索还可以用于找到两个节点的最短路径问题。

社区发现是近年来科学知识扩散的研究热点之一，可用于知识团体发现等。Louvain 和标签传播（Label propagation）是两种实际应用较多的社区发现算法。

Louvain 算法是一种基于多层次（逐轮启发式迭代）对模块度进行优化的算法。其核心模块度是评估一个社区网络划分好坏的度量方法，物理含义是社区内节点的连边数与随机情况下的边数之差，取值范围是[-1/2,1)，它在 Louvain 算法中被用来当作一个优化函数（目标函数），即将结点加入它的某个邻居所在的社区中，如果能够提升当前社区结构的模块度，则说明迭代优化可接受。

标签传播算法是一种基于图的半监督学习方法，其基本思路是用已标记节点的标签信息去预测未标记节点的标签信息。利用样本间的关系建立关系完全图模型，在完全图中，节点包括已标注和未标注数据，其边表示两个节点的相似度，节点的标签按相似度传递给其他节点。标签数据就像是一个源头，可以对无标签数据进行标注，节点的相似度越大，标签越容易传播。与 louvain 算法相比标签传播算法更为简单快速，缺点是每次迭代结果不稳定，准确率不高。

（2）数理统计与机器学习相关算法

科学知识扩散能力测度相关研究大多涉及数理统计与机器学习，尤其是回归和聚类两种方法，在相关研究中占有大量比重。

回归分析是一种预测性的建模技术，它能够表明自变量和因变量之间的显著

关系和多个自变量对一个因变量的影响强度。选择合适的回归算法主要涉及自变量的个数、因变量的类型以及回归线的形状三个方面。常用的回归算法包括：

线性回归（Linear regression），要求因变量连续，自变量为连续或离散，使用最佳拟合曲线在因变量（Y）和一个或多个自变量（X）之间建立关系。其方程式表示为 Y=a+b*X + e，其中 a 表示截距，b 表示直线的斜率，e 是误差项。根据给定的预测变量可以预测目标变量的值。其中最佳拟合曲线通常通过最小二乘法最小化平方误差取得。

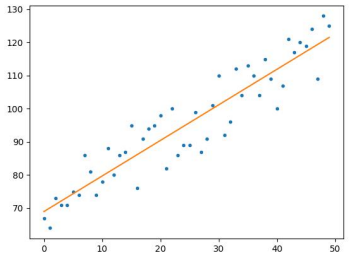

图 1 线性回归拟合曲线

逻辑回归(logistic regression)用于计算“事件=Success”和“事件=Failure”的概率，当因变量的类型属于二元（1 / 0，真/假，是/否）变量时使用。其方程式表示为 $\text{logit}(p)=a+b_1x_1+...+b_nx_n$ ，其中 a=常数项，表示自变量 x 取值为 0 时，比数（Y=1 与 Y=0 的概率之比）的自然对数值，p 表示概率。参数的选择通过观测样本的极大似然估计值进行选择。

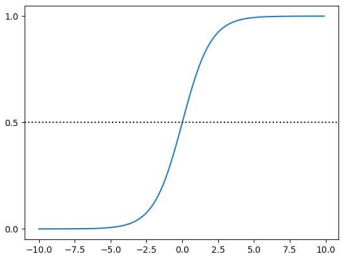

图 2 逻辑回归拟合曲线

多项式回归（polynomial regression）与线性回归相似，但其自变量的指数大于 1，其方程式表示为 $y=a+b_1x+b_2x^2+...+b_nx^n$ ，最佳拟合线不是直线而是曲线。

逐步回归（step regression）通过观察统计的值，如 R-square，t-stats 和 AIC 指标，来识别重要的变量。逐步回归通过同时添加/删除基于指定标准的协变量来拟合模型。常用的逐步回归方法包括向前选择（从模型中最显著的预测开始，然后为每一步添加变量）和向后剔除法（模型的所有预测同时开始，然后在每一步消除最小显着性的变量），这种建模技术的目的是使用最少的预测变量数来最大化预测能力，是处理高维数据集的方法之一。

其他的回归算法还有用于存在多重共线性（自变量高度相关）数据的岭回归

（ridge regression）、套索回归（lasso regression）等。

聚类(Clustering)是按照某个特定标准(如距离)把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。当前已发展出多种不同种类的聚类算法。

划分式聚类算法(Partition-based clustering)需要事先指定簇类的数目或者聚类中心，通过反复迭代。经典划分式聚类算法如 K-means 等。

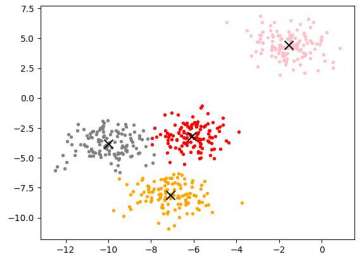

图 3 K-means 聚类算法示例

划分式聚类算法的不足之处在于相似簇的归并会放大一个簇内的相似性误差。层次聚类算法 (hierarchical clustering)是另一种典型的聚类算法，解决了上述问题。它将数据集划分为一层一层的簇，后面一层生成的簇基于前面一层的结果。Agglomerative 是实际使用中应用较多的算法。

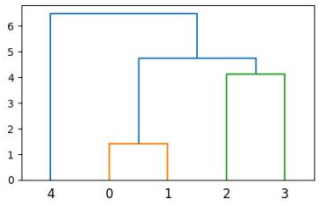

图 4 Agglomerative 聚类算法示例

此外还有基于密度的聚类算法(Density-based clustering)如 OPTICS 算法，以及一些其他聚类算法如核聚类、谱聚类等。

（3）深度学习与自然语言处理相关算法

应用深度学习与自然语言处理方法进行文本挖掘是近年来科学知识扩散研究的新趋势，应用较多的算法有语义词向量的计算算法（词嵌入），以 Word2vec 为代表，基于文本相似度度量知识扩散；使用主题模型如 LDA 等进行主题知识单元提取；以及深度学习在文本挖掘方面的应用如经典的循环神经网络（Recurrent Neural Network，RNN）和长短时记忆神经网络（Long Short Term Memory，LSTM）以及 BERT 等。

在使用独热编码等对自然语言进行建模的过程中，会出现维数灾难、词语相似性、模型泛化能力以及模型性能等问题。寻找上述问题的解决方案是推动统计语言模型不断发展的内在动力。在对统计语言模型进行研究的背景下，Word2vec 产生。它是一种非监督算法，可以根据给定的语料库，通过优化后的训练模型快

速有效地将一个词语表达成向量形式，为自然语言处理领域的应用研究提供了新的工具。Word2vec 依赖 skip-grams 或连续词袋模型（CBOW）来建立词嵌入。其中连续词袋模型通过上下文来预测当前值，而 skip-grams 则通过当前词预测上下文。

隐狄利克雷算法（Latent Dirichlet Allocation，LDA）用来推测文档的主题分布。它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题分布后，根据主题分布进行主题聚类或文本分类。所谓主题可以被定义为“语料库中具有相同词境的词的集合模式”，比如说，主题模型可以将 “农场”，“玉米”，“小麦” 集合成 “农业”主题等。LDA 的数学推导较为复杂，可以简单理解为，它是一种生成式模型，包含词、主题 和文档三层结构，认为一篇文章的每个词都是通过“文章以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”的过程得到。文档到主题和主题到词都服从多项式分布。

循环神经网络（Rnn）是神经网络的一种，尤其适用于处理如文本等序列化数据。RNN 的特征在于，对于每个 RNN 神经元，其参数始终共享，即对于文本序列，任何一个输入都经过相同的处理，得到一个输出。在传统的全连接神经网络的结构中，神经元之间互不影响，并没有直接联系，神经元与神经元之间相互独立。而在 RNN 结构中，隐藏层的神经元通过一个隐藏状态所相连，其示意图如下所示：

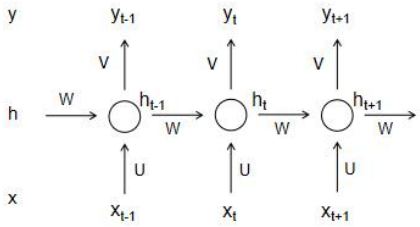

图 5 RNN 结构示意图

由于 RNN 单元在面对长序列数据时，容易产生梯度弥散，使得 RNN 只具备短期记忆，即 RNN 面对长序列数据，仅可获取较近的序列的信息，而对较早期的序列不具备记忆功能，从而丢失信息。为此，为解决该类问题，LSTM 结构被提出，其核心关键在于门机制的设立（遗忘门、输入门、输出门）以及细胞状态，使得其能保存长距离信息。LSTM 示意图如下所示：

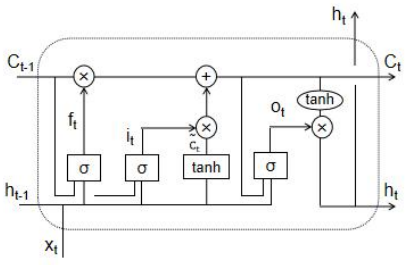

图 6 LSTM 结构示意图

BERT 的全称为 Bidirectional Encoder Representation from Transformers，是一个预训练的语言表征模型。它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的 masked language model（MLM），以致能生成深度的双向语言表征。该模型有以下主要优点：采用 MLM 对双向的 Transformers 进行预训练，以生成深层的双向语言表征；预训练后，只需要添加一个额外的输出层进行 fine-tune，就可以在各种各样的下游任务中取得 state-of-the-art 的表现。在这过程中并不需要对 BERT 进行任务特定的结构修改。BERT 已在各种自然语言处理相关的任务中得到应用。

（4）系统动力学相关算法

系统动力学将科学知识扩散置于系统的形式加以考察。在确定系统的边界后，用计算机直接建立系统模型并通过模拟计算了解系统随时间变化的行为和特性。系统动力学与因果关系模型等直接相关，其主要建模过程包括因果回路图：分析系统中的要素，界定好箭头及各回路的极性；存量流量图：找出水平变量、辅助变量、常量等，系统庞大时可借助影子变量将系统拆分为几个子系统；系统动力学方程：包括设置方程、单位、初始值、时长、开始和结束的时间等，界定系统的界限，做出假设等。

1.4.2. 扩散测度的分析工具

科学知识扩散能力的分析测度是一项系统性工作，主要的分析工具涵盖数据存储相关的数据库与查询语言、计算机辅助计算的计算框架、网络与计量相关的可视化、系统动力学相关的模拟仿真等诸多方面。

（1）数据存储

探究科学知识的扩散需要收集相关数据并进行存储。用于存储数据的数据库包括关系型数据库和非关系型数据两种主要类别。

关系型数据库，是指采用了关系模型来组织数据的数据库，以行和列的形式存储数据，一系列的行和列被称为表。通过设置不同形式的外键来体现表和表的不同关系，一组表组成了数据库。较为常用的关系型数据库包括 Oracle Database、SQL Server 、MySQL 等。

非关系型数据库，又称 NoSQL 数据库。相较关系型数据库而言，非关系型数据库更易扩展，具有更高的读写性能，因而近年来更多地被用于科学知识相关的数据存储。常见的非关系型数据库有 Neo4j、MongoDB、Jena 等。

（2）数理统计与分析

在获取相关科学知识数据后，可对其进行数理统计与分析以实现科学知识扩散能力的定量测度。当前常用的统计分析软件有 SPSS（Statistical Product

Service Solutions）和 stata 等，它们被用于统计学分析运算、数据挖掘、预测分析和决策支持任务。相关的统计分析过程包括描述性统计、均值比较、一般线性模型、相关分析、回归分析、对数线性模型、聚类分析、数据简化、生存分析、时间序列分析、多重响应等。具有专门的绘图系统，可以根据数据绘制各种图形。

（3）数据可视化

数据可视化借助图形化手段，能够直观展示复杂信息，帮助理解数据。其中 python 是当前常用的数据分析语言，javascript 是常用的 web 页面脚本开发语言。　基于 python 的绘图库包括 Matplotlib 、Seaborn、Plotly 等。其中 Seaborn 在 Matplotlib 的基础上进行了更高级的 API 封装，能高度兼容各种数据结构与统计模式；plotly 则是一个基于 javascript 的绘图库，但可以通过 python 编程进行使用，并且可以与 Web 无缝集成；基于 javascript 的绘图库包括 Echarts.js、D3 等。 它们遵循现有的 Web 标准，将强大的可视化组件和数据驱动的 DOM 操作方法结合在一起。

（4）机器学习与深度学习

随着计算机技术的发展，当前研究者开始通过机器学习、深度学习（主要是自然语言处理）等数据挖掘手段对科学知识数据作出进一步分析。较为常用的 python 机器学习包和当前应用较广泛的深度学习框架有：

NLTK，全称 Natural Language Toolkit，自然语言处理工具包，是 NLP 研究领域常用的一个 Python 库。其功能包括字符串处理、搭配发现、词性标识、分类、分块等，也可用于实现分析器，wordnet 查看器，聊天机器人等简单应用。

Gensim（generate similarity）是一个简单高效的自然语言处理 Python 库，用于抽取文档的语义主题（semantic topics），内置的算法包括 Word2Vec，FastText，潜在语义分析（Latent Semantic Analysis，LSA），潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）等。

Scikit-learn 是一个专门面向机器学习的 python 科学计算工具包，基本功能包括分类，回归，聚类，数据降维，模型选择和数据预处理等。

就深度学习而言，Tensorflow 和 Pytorch 是当前使用最广泛的两个计算框架。TensorFlow 是谷歌发布的深度学习开源的计算框架，其中 Tensor 翻译成“张量”，是一种多维数组的数据结构；Flow 翻译成“流”，是计算模型，描述的是张量之间通过计算而转换的过程。计算图是由节点和边组成的，而 TensorFlow 是一个通过计算图的形式表述计算的编程系统，每一个计算都是计算图上的一个节点，节点之间的边描述了计算之间的关系。PyTorch 是一个开源的 Python 机器学习库，基于 Torch，用于自然语言处理等应用程序，包含自动求导系统的深度神经网络。

（5）网络分析与可视化

通过网络或者科学知识图谱对科学进行表征和测度也是常用的研究方法。相关网络可视化和分析软件包括 Citespace、VOSviewer 和 Gephi 等。

CiteSpace 是一款引文可视化分析软件，通过可视化的手段来呈现科学知识的结构、规律和分布情况，其分析得到的可视化图形称为“科学知识图谱”。使用 CiteSpace 可以进行共被引、共词、突现、聚类等多种方面的分析，也可以挖掘知识基础、定位关键文献、获取学科结构、探索学科前沿等。与 CiteSpace 类似，VOSviewer 也是一款文献可视化工具，可生成任何类型的文本地图，可以对文献进行多方面的分析。二者共同的不足之处在于对输入的数据格式有一定限制，只接受一定学术数据库的来源数据，当前可接受的数据来源包括 WoS, Scopus, Derwent, CNKI, CSSCI, CSCD 等知名数据库。

Gephi 是一款网络分析领域的数据可视化处理软件，支持模块化扩展及插件开发。其主要功能包括网络布局、网络统计（即通过不同的统计算法计算网络属性，发现网络节点和边、网络整体和小团体的相关特性）；网络滤波，通过用户设定的规则对网络中的节点或边进行筛选，从而更加精准的探索和分析网络；网络可视化等，相比前两个软件功能更加自由。

（6）系统动力学仿真

由于科学知识及相关行为如知识扩散等实际构成了非线性复杂系统，因此通过系统动力学对系统的演变和相关影响因素进行分析成为科学知识扩散的一种重要研究方法。相关系统动力学仿真软件有 STELLA 和 Vensim 等。主要功能包括映射和建模，即用直观的基于图标的图形界面简化模型的构建，以图形化的各式箭头记号连接各式变量记号，并将各变量之间的关系以适当方式写入模型，记录各变量之间的因果关系。同时对模型提供多种分析方法：包括结构分析工具和数据集分析工具，以及模型真实性检验。

1.5.本章小结

科学知识的扩散有利于技术进步与经济发展，也有利于社会文化观念的塑造。所谓科学知识扩散能力即代表知识基于不同载体在不同知识发出和接收个体之间的传播能力。狭义的理解是对科学知识扩散能力的测度指对已有知识扩散情况的度量，如扩散的广度、强度、速度等。然而对于科学知识扩散能力的测度不能仅限于此，影响科学知识扩散情况的条件因素研究、预测其未来发展情况等也是科学知识扩散能力测度的目标之一，且具有更强的现实意义。

相应地，科学知识扩散能力的数值指标仅代表知识扩散能力狭义的一方面，广义而言更为重要的是科学知识扩散能力测度的研究方法体系。伴随技术的发展与进步、数据驱动和实证主义研究范式的发展、对科学知识扩散的交叉学科研究属性的认识，新的研究范式开始产生，基于数学、物理、计算机与情报学等多学科方法的移植和创新实现对科学知识扩散能力的定量方法研究逐渐成为主流，最

显著的代表就是复杂网络、自然语言处理、系统动力学等在科学知识扩散能力研究中的应用。针对特定研究问题，相关算法不断得到发展，分析工具也不断推陈出新，这就要求研究人员掌握多学科知识、跟进学科研究前沿，能够快速学习和掌握必要的软件工具。

最后，当下科学知识扩散的载体和途径多样，正如传统研究范式——引文网络不能代表和概括当今的科学知识扩散研究，如今的科学知识扩散能力测度的维度、属性与方法也难以涵盖未来科学知识扩散的不同情况。期望研究者们能在对已有知识融会贯通的基础上，设计发展新的研究方法和研究框架以适应变化的科学知识扩散状况，共同推动相关学科的交叉融合以及科学知识扩散研究的进步。