专利引文分析的理论与实践

The Theory and Practice of Patent Citation Analysis

杨中楷　梁永霞○著

科学出版社

大连理工大学人文与社会科学学部学术著作出版资助项目

专利引文分析的理论与实践

The Theory and Practice of Patent Citation Analysis

杨中楷　梁永霞○著

科学出版社

北京

图书在版编目（CIP）数据

专利引文分析的理论与实践 / 杨中楷，梁永霞著. —北京：科学出版社，2017.9
ISBN 978-7-03-054368-4

Ⅰ. ①专… Ⅱ. ①杨… ②梁… Ⅲ. ①专利-引文分析-研究 Ⅳ. ①G306.0

中国版本图书馆 CIP 数据核字（2017）第 218918 号

责任编辑：邹 聪 陈会迎 / 责任校对：韩 杨
责任印制：张欣秀 / 封面设计：有道文化
联系电话：010-64035853
电子邮箱：houjunlin@mail.sciencep.com

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
北京虎彩印刷科技有限公司印刷
科学出版社发行 各地新华书店经销
*
2017 年 9 月第 一 版 开本：720×1000 B5
2017 年 9 月第一次印刷 印张：13 1/4 插页：2
字数：210 000
定价：68.00 元
（如有印装质量问题，我社负责调换）

序

2000 年前后，由我的博士生导师刘则渊教授牵头，大连理工大学科学学与科技管理研究所开始重点推进科学计量学研究。在经历了开创期的诸多困难之后，刘老师带领着年轻的教师和博士生们在科学计量学的道路上不断前进，取得了令业界瞩目的科研业绩。2005 年，刘老师与大连理工大学海天学者特聘教授克雷奇默博士一起创办 WISE 实验室。WISE 是网络计量学（webometrics）、信息计量学（informetrics）、科学计量学（scientometrics）和经济计量学（econometrics）的缩写。WISE 实验室一经成立，便在国内科学计量学界崭露头角，也吸引了国际学术界的关注，逐渐成为全球科学计量学研究的重要基地。2007 年，美国德雷塞尔大学教授、CiteSpace 软件的创始人陈超美受聘长江学者讲座教授，带来了信息可视化和知识可视化的新思想，大连理工大学的科学计量学研究也因此再上新台阶。

大连理工大学科学计量学研究产生重大变革的历史时期，正是我从一名博士生转变为一名青年教师的过渡时期，也是我个人研究方向从模糊不定到清晰明确的关键时期。在做博士学位论文的时候，作为对科学计量学研究的拓展，我选择了专利计量作为选题方向。在研究过程中，我借鉴科学计量学的研究方法，利用海量专利文献开展研究，做了一些专利计量方面的开创性工作。我也注重利用一些可视化工具软件，推进专利计量的可视化研究，在专利引文分析及其可视化方面取得了一些进展和突破。博士毕业留校任教之后，我遵从刘老师的建议，继续开展专利计量和专利引文研究，发表了一系列学术论文。其中，2010 年发表在《科研管理》上的论文《专利引用过程中的知识活动探析》被中国科学技

术信息研究所遴选为F5000项目入选论文。2011年发表在《科学学研究》上的论文《基于专利引文网络的技术轨道识别研究——以太阳能光伏电池板领域为例》被多次引用，成为技术轨道识别研究的重要论文之一。

在此期间，我指导了两名硕士研究生从事专利引文分析研究。硕士生刘倩楠的学位论文《基于专利引文网络的技术演进路径识别研究》较早地运用专利引文网络进行了技术演进路径研究，是国内专利引文网络研究的开创研究之一。硕士生于霜的学位论文《基于专利引文网络的空间关系可视化研究》利用百万数量级的专利引文数据，刻画了基于专利引文的地理单位和技术领域的空间联系，同样是国内专利引文网络研究的开创研究。我的硕士生沈露威、刘佳、黄颖、徐梦真、韩爽等虽然没有将专利引文分析作为学位论文选题，但也在此领域做出了非常优秀的研究成果，获得了校内外各种学术荣誉。

在从事专利计量和专利引文分析研究的过程中，我发现当前专利引文分析方法和应用研究较多，但与专利引文分析基础和机理相关的研究较少。基于此，我认为应该从源头出发，理顺专利引文分析的研究脉络，构建从理论、方法到实践的知识全链条。此想法得到了《中国科技期刊研究》编辑部主任梁永霞博士的响应。梁博士既是我的师妹，也是我多年的学术合作伙伴，在她的博士学位论文《引文分析学的知识计量研究》中，已经将专利引文分析作为引文分析学的一个分支进行重点考察。我们尝试以刘老师提出的知识流动过程中知识的扩散和重组理论为基调，再结合我个人和我的学生们的专利引文分析的应用研究成果，对专利引文分析进行了专门的、系统的论述。经过数月的文字工作，终于能够呈现给读者们这部专注于专利引文分析的作品。

本书本着从一般性到特殊性、从普遍性到个别性的原则，分6章对专利引文分析的基本概念、理论和方法体系进行较为系统的阐述。第1章主要阐述引文分析的基本理论问题，重点阐述知识流动理论这一贯穿引文分析始终的重要理论基础。第2章则基于知识流动理论阐述专利引文分析的相关概念，分析专利引文网络中知识活动的基本原理，提出专利引文分析的研究框架。第1章和第2章属于理论研究范畴，目的是为后面的实证研究奠定基础。第3章和第4章利用海量专利引文数据，分

别揭示地理空间和技术空间中的知识流动情况，形象地展现专利引文分析对知识扩散轨迹的追踪作用。第 5 章基于专利进化树的原理提供一种识别专利引文网络中技术演化轨道的方法，生动简洁地展示出复杂引文网络中的技术进化过程，并用两个案例进行验证。第 6 章呈现专利引文分析用于定量评价的一面，利用一系列专利引文相关指标对技术发展的特征进行测度与展示。

在本书编写和出版过程中，我的在读研究生高霞、王雪莹、侍晓宇、刘倍言、孙昕和准研究生苏英协助做了不少工作，在此表示感谢。同时要感谢科学出版社科学人文分社的侯俊琳分社长、邹聪编辑为本书出版付出的辛勤努力，也要感谢大连理工大学人文与社会科学学部为本书出版提供的经费支持。

虽然目前国内专门研究专利引文分析的著作较少，但是关于专利计量研究的成果已不在少数。作者中不但有邱均平教授、黄鲁成教授等知名专家，也有文庭孝教授、王贤文教授等青年才俊。希望能够通过本书的出版，就教于业内同行，也希望借此抛砖引玉，为推动我国专利计量研究领域的发展略尽微薄之力。

杨中楷

2017 年 8 月于大连理工大学科技园

目　录

第 1 章　引文分析的基本理论问题

1.1　引文分析的相关概念

引文分析（citation analysis），是一种对文献引证与被引证关系进行分析的活动和方法，也是包含对引文关系进行分析的原理、方法、应用在内的一门学科。引文分析是基于文献间的联系而产生的一种分析方法。具体来说，文献体系中文献之间并不是孤立的，而是相互联系的。文献的相互关系突出地表现在文献的相互引用方面。一篇文献在编写过程中一般都需要参考有关文献。在文献发表时，作者往往采用尾注或脚注等形式列出其“参考文献”或“引用书目”。一个“引文”是指一篇参考文献，进行引用的是引用（citing）文献，接受引用的是被引（cited）文献。普赖斯在论及引证及被引证关系时提出：每一篇被引文献，对于引证者（文献作者）来说，就是有了一篇参考文献，而对于被引证者来说，则是有了一篇引证文献（引文）。

一篇文献既可以是施引文献，也可以是被引文献。我们谈到引文时，可以站在两个角度：一是站在施引文献的角度，那么引文就是其参考文献；二是站在被引文献的角度，引文就是其本身。引文是有方向的，施引文献的时间一般比引文要晚，不可能倒过来引用。

文献在被引时，不一定是全部内容被引，因此，可以把一篇文献中被引的部分称为知识单元，那么知识单元就有生产单元和储存单元之分。如果文献 A 中含有使用并描述文献 B 的书目注释，那么文献 A 就含有文献 B 的参考文献，而文献 B 具有来自文献 A 的引文。在上述的过程中，A 被称为引用文献，而 B 被称为被引用文献。按照期刊间引用关系的概念——知识生产单元和知识存储单元（Zinkhan and Leigh，1999），

我们也称 A 为知识存储单元，B 为知识生产单元（埃格希和鲁索，1992）。知识从 B 流向 A，如图 1.1 所示，意味着引用是个动态的过程。

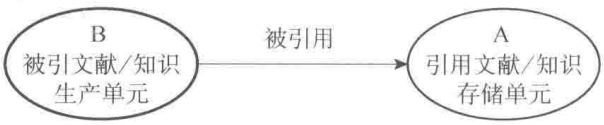

图 1.1　引用过程对应的概念

当引文网络中的文献不是很多（少于几百个）时，用一张引文图就可以形象地表达文献之间的引用关系。箭头从代表 d_i 的一端指向代表 d_j 的一端时，来自某一馆藏的文献就形成一张有向图，这张图就称为“引文图”或“引文网络”（图 1.2）。

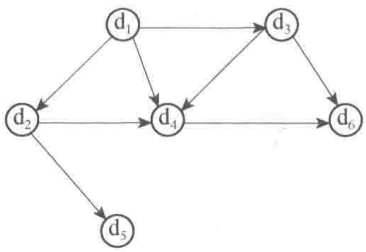

图 1.2　引文图

利用引文图表达引文关系的一个优点是比较明晰、清楚，引文关系一目了然。但是，如果引文图中涉及的文献很多（数百个以上），那么图形就变得相当复杂，很难分析出关系的结构，这是引文图的一个缺点。在这种情况下，最好利用矩阵方法来表达关系网络（尹丽春，2006），引文分析方法也真正有了用武之地。

引用过程是单个的、个体的，是慢慢积累起来的，而引文分析的过程包括对引用过程及海量数据的分析。引文网络是一个知识生产和传播的复杂系统，个人和单个文献的作用在网络中已经逐步淡化，仅仅依赖于同行评议和单纯地分析个体文献无法真实地反映整个网络的状态。只有通过数学手段将网络的整体结构绘制出来，人们才可能从全局着手做出全面而正确的判断。超大规模引文网络的形成迫切需要科学工作者提出有效的手段对其进行研究（尹丽春，2006）。

总的来说，引文分析就是利用各种数学及统计学的方法进行比较、

归纳、抽象、概括等的逻辑方法，对科学期刊、文献、著者等分析对象的引用和被引用现象进行分析，以揭示其数量特征和内在规律的一种信息计量研究方法（邱均平，2007）。

关于引文分析，有两个概念不得不提。首先是凯斯勒在 1963 年提出了文献耦合（bibliographic coupling）的概念。文献耦合是指引证文献通过其参考文献（被引证文献）建立的耦合关系。具体来说，如果 A 和 B 两篇文献共同引证了一篇或多篇参考文献，或者说它们有一篇或多篇同样的参考文献，则称 A 和 B 两篇文献具有引文上的耦合关系。

另一个对应的概念是文献共引（bibliographic co-citation），也称同引、同被引、共被引，是由美国的斯莫尔和俄罗斯的玛莎科娃在 1973 年分别独立提出来的，就是指两篇或多篇文献同时被后来的一篇或多篇文献所引证，则称这两篇文献（被引证文献）具有“同被引”关系。图 1.3（a）展示了 A、B 两个文献同时被文献 a 引用的状况，图 1.3（b）则展示出 A、B 两篇文献同时引用文献 a 的情况。

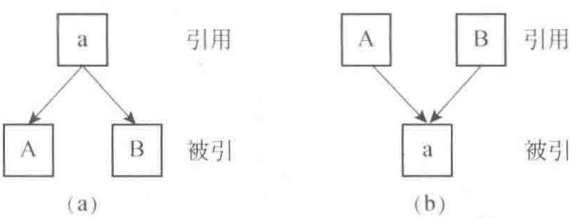

图 1.3　文献的共引与耦合

1.2　引文分析的整体发展脉络

科学知识可视化图谱是在信息技术的推动下发展出来的一个新领域，当前已经成为科学计量学的一个新热点。陈悦和刘则渊（2005）认为科学知识图谱是显示科学知识的发展进程与结构关系的一种图形，它是揭示科学知识及其活动规律的科学计量学从数学表达转向图形表达的产物，是显示科学知识地理分布的知识地图转向以图谱展现知识结构关系与演进规律的结果。为揭示引文分析领域的历史图景，选取 CiteSpace 绘制了引文分析领域演进知识图谱，从而清晰地看出引文分析学形成和

发展的脉络及演进趋势。

研究所用的数据来源于美国科学情报研究所创建的 Web of Science 数据库。以“citation analysis”为检索词在科学引文索引（Science Citation Index，SCI）和社会科学引文索引（Social Sciences Citation Index，SSCI）中联合检索了 1974～2008 年的文献记录。在数据下载的过程中，我们选择“Article”，共检索到 1906 篇文献，其中共包含引文 65 426 条。对得到数据的引文进行整理和标准化，力图使引文数据准确。

利用 CiteSpace 软件，输入题录数据，选择“cluster”分析，同时设置阈值为(3, 2, 15)、(4, 3, 19)、(4, 3, 20)，网络节点选为参考文献(reference)，来源选为文献标题（title）、摘要（abstract）、关键词（descriptor）和标识符（identifiers），术语选择为无（none），修剪（pruning）项选择最小生成树（minimum spanning tree）、修剪分段的网络（pruning sliced networks）、修剪混合网络（pruning merged the network）得到引文分析的发展趋势网络，其中共有节点 86 个，连线 114 条（图 1.4）。

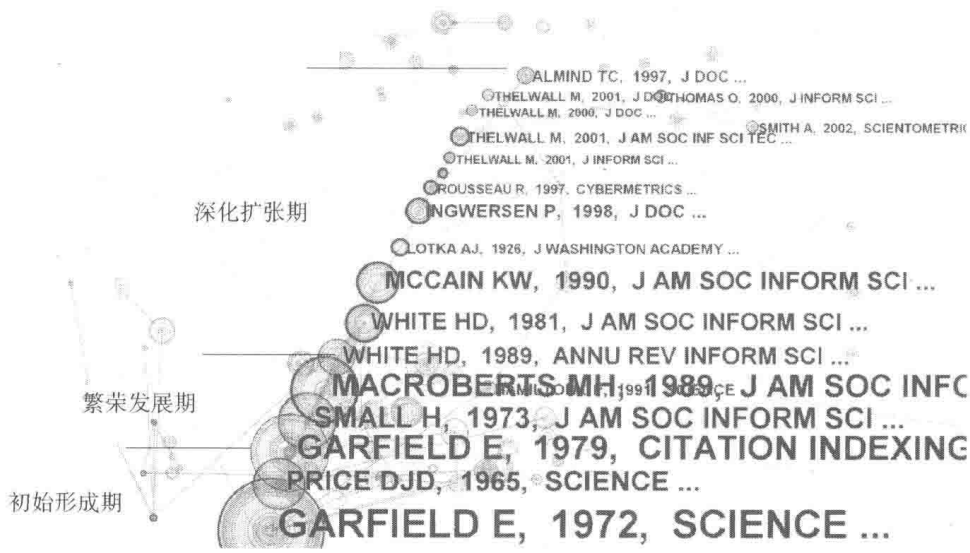

图 1.4　引文分析的最小生成树图

由图 1.4 可以看到，引文分析领域大致可以分为三个时期：初始形成期、繁荣发展期、深化扩张期。初始形成期中可看出关键人物有加菲尔德和普赖斯，他们二人开创了引文分析的先河，是引文分析学的奠基人。发展繁荣期中的重要人物有斯莫尔和麦克罗伯特，他们二人发展了引文

分析，其中斯莫尔提出了著名的共引理论和方法，而麦克罗伯特则思考了引文分析存在的问题。从 20 世纪 80 年代起，引文分析进入了深化扩张期。在共引理论的基础上，引文分析的可视化有了较大的发展，重要的人物有怀特、麦肯恩和陈超美等。90 年代中后期，随着互联网的快速发展，网络引文分析也成为引文分析的热点，其代表人物有英格沃森、塞沃尔与鲁索等。当然，由于阈值的设置，这张图谱只能大致反映引文分析领域最重要的人物和著作。

为了能够更加形象化地展示引文分析领域的形成和发展，更加清楚地看到引文分析发展的脉络，仍旧利用 CiteSpace 软件，输入题录数据，选择“cluster”分析，同时设置阈值为（3, 2, 15）、（4, 3, 20）、（3, 3, 20），网络节点选为参考文献（reference），来源选为文献标题（title）、摘要（abstract）、关键词（descriptor）和标识符（identifiers），术语选择为无（none），修剪（pruning）项选择最小生成树（minimum spanning tree）、修剪分段的网络（pruning sliced networks）、修剪混合网络（pruning merged the network），并且利用 time-zone，得到引文分析的演进网络，其中共有节点 202 个，连线 2033 条（图 1.5）。

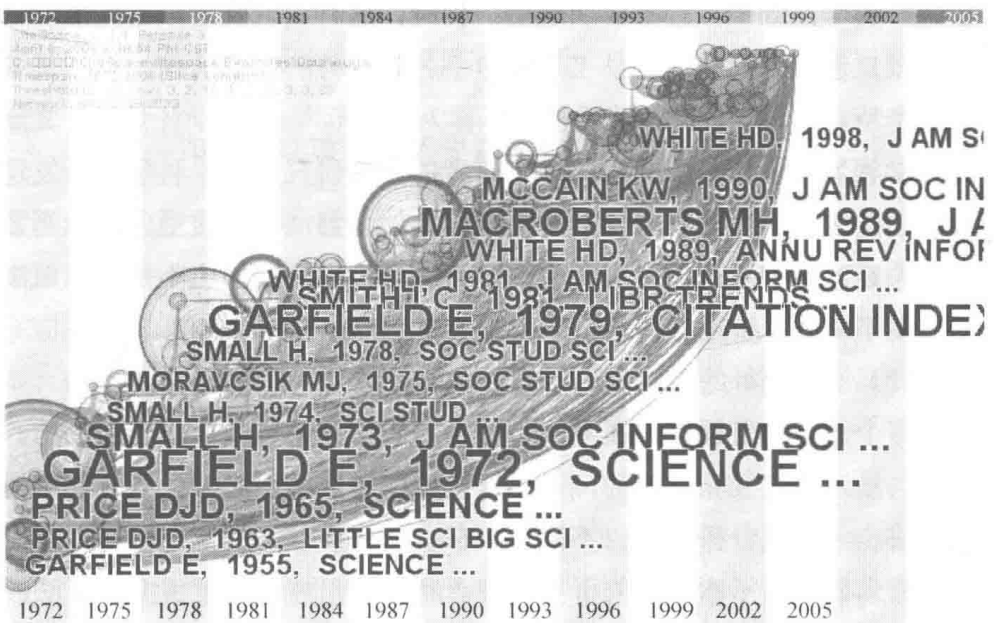

图 1.5　引文分析研究发展趋势图（文后附彩图）

由图 1.5 可以清楚地看到引文分析的发展脉络及更多的代表人物。我们可以看到加菲尔德和普赖斯都对引文分析学的理论做出了开创性的工作。因此，可以把他们的理论和方法认为是引文分析学的奠基时期的重要贡献。在加菲尔德和普赖斯关于科学引文网络的思想的基础上，著名科学计量学家斯莫尔也为引文分析的进一步发展做出了不可磨灭的贡献。共引分析也是引文分析独特的分析方法，迄今为止，共引分析仍旧是引文分析的主流方法。

通过对引文分析领域发展的研究和梳理，可以看到引文分析领域的发展有以下的规律。

（1）引文分析成为科学计量学、文献计量学的范式

库恩（2003）指出，“常规科学”是指严格根据一种或多种已有科学成就所进行的科学研究，某一科学共同体承认这些成就就是一定时期内进一步开展活动的基础。研究工作可以明明白白地从一套规则中引出来，但范式却比任何一套这样的规则都要更为优先、更为适合、更加完整。“一种成就”构成了范式。自从加菲尔德、普赖斯、斯莫尔等的著作出版后，这些著作的成就足以空前地把一批坚定的拥护者吸引过来，使他们不再去进行科学活动中各种形式的竞争。同时，这种成就又足以毫无限制地为一批重新组合起来的科学工作者留下各种有待解决的问题。凡是具备这两个特点的科学成就，此后被库恩称为“规范”。在科学计量学、文献计量学领域，引文分析成为科学工作者的主要研究范式。科学的新发现或者新发明对于研究者来说并不是刚开始就全部能够被接受的，也需要一个不断适应和成长的过程。引文分析这个领域的发展也不是一帆风顺的，是不断解答质疑者、不断改进，从而完善和成熟起来的。

（2）引文分析的发展受到了新技术的促进

每个学科的发展都是在吸取其他学科的精华的过程中成长起来的。科学与技术也是互相促进的，一个新技术的发展会引起相关科学领域的巨大进步，引文分析领域也不例外。在引文分析发展的三个时期，第一是加菲尔德吸取了谢泼德的引文的理念和技术，才促成了引文索引法的诞生，弥补了主题索引法的不足，能够更加准确快速地找到研究者所需的文献。第二就是计算机技术的大发展为引文分析的可视化提供了很好

的土壤和平台，从过去的手工绘图发展为机器绘图。第三就是互联网的兴起和发展，促进了知识的快速流动，也为海量的数据库提供了可能，为引文分析提供了绝好的网络环境，能够更加及时地发现引用关系。同时互联网的出现对引文分析研究从方法和内容上都提出了新的挑战，传统的引文分析能否和网络计量无缝连接，这是对引文分析领域研究者的新考验。

刘则渊等（2008）在《科学知识图谱：方法与应用》一书的导言中，明确指出："从普赖斯、加菲尔德到斯莫尔，已确立起日臻完备的引文分析理论与方法，构成科学计量学的基础与主流，在一定意义上也可以说在科学计量学中已形成一门成熟的分支学科——引文分析学。"引文分析学是以文献、信息、科学、技术、网络等领域的引证和被引证关系为对象，以引文之间知识联系与流动过程为基础，运用数学、社会学、心理学、图形学、计算机学等现代方法和手段，研究引文活动的规律及其应用的一门交叉学科。引文分析学与文献计量学、科学计量学、信息计量学、网络计量学、知识计量学等有密切的关系，其共同基础都是知识流动理论。

1.3　引文过程中的知识流动理论

1.3.1　知识流动理论的历史源流

波普尔在《客观知识——一个进化论的研究》一书中明确提出科学知识增长图式："P1→TT→EE→P2"，即"问题→试探性理论→排除错误→新问题"。科学知识的增长过程，就是各种不同的假说（理论）互相竞争、自然选择的过程，这一过程与生物进化的过程十分相似。波普尔区分三个适应层次：遗传适应、适应性行为学习和科学发现。他认为一切进化都含有一种基本的承袭结构，它在不同的领域表现为有机体的基因结构、基本行为形式和公认的理论（周超，2004）。波普尔（2005）的世界 3 理论主要致力于为世界 3 的客观性、自主性和实在性作辩护。他一再强调，客观知识是由说出、写出、印出的各种陈述组成的，如科学知识是由问题、问题境况、假说、科学理论、论据等组成的。客观知识包

括思想内容及语言所表述的理论内容，它们出现在杂志、书本、图书馆等一定的环境之中。知识的客观意义可与蜜蜂酿的蜜相类比。人在适应自然界的过程中，获得知识是最主动的，也许甚至比获取食物还主动。自然界的任何信息都不会自行从环境中流入人的头脑中，人只有像寻求食物一样主动地探索自然界并从中汲取信息，才能从自然界获得有用的信息。在波普尔看来，对科学知识积累最有意义的事件是证伪旧理论，而不是证实新理论（赵敦华，2006）。

库恩提出的科学知识进化的“范式”模式认为科学知识依照“前科学（没有范式）→常规科学（建立范式）→科学革命（范式动摇）→新常规科学（建立新范式）”的知识增长模式或者简单地是按范式→范式的过程进化，与波普尔模式有着不同的侧重点。他指出了问题（即反常、危机）在知识进化中的航标作用。“危机的意义就在于，它可以指示更换工具的时机已经到来”（库恩，2003），危机给科学家带来了批判精神和创造精神，各种竞争的理论层出不穷，经历了百家争鸣的新范式选择后，一切危机都随着新范式的出现及其被接受而宣告结束，从而确立了新的常规科学。波普尔的模式侧重于从知识创造过程→思维过程的角度讨论科学知识的进化，而库恩模式则主要是从思维结果→科学理论的互相更替来谈科学知识进化的。因此，它们之间不应该互相对峙而是互补的。

实际上，知识本身就隐含了创新的特质，著名的物理学家、科学学创始人之一贝尔纳（1982）曾指出：“科学远远不仅是许多已知的事实、定律和理论的总汇，而是许多新事实、新定律和新理论的继续不断的发见。它所批评的，以及常常摧毁的东西，同它所建造的东西一样多。”由此可见，科学知识的进化是一个对旧的科学观念的否定的过程，它必然导致科学理论的焕然一新，必然要伴随自然观、方法论和思维方式的全面变革。

迄今为止，人们对知识进化的研究成果，形成了不同的学说和模式，归纳起来主要有：“证伪主义模式”（拉卡托斯，1986）、“历史主义模式”（施良方，1994）、“研究纲领模式”（拉卡托斯，1986）、“信息加工论”（施良方，1994）、“仿生物进化论”[包括“自然选择论”、“知识基因论”（刘

植惠，1999)]等。这些知识进化学说和模式都是基于文献的知识的进化，或者说这些进化学说和模式的建立都离不开文献。

1.3.2 引文分析中知识流动的基本观点

1）知识进化的显著特征是知识的继承的“遗传”性和发展的“变异”性。继承是知识进化发展的基础，发展是知识进化的标志和目的。显然，这种“继承”和“发展”都离不开对文献的利用（马恒通，2005）。文本中的客观知识成为人脑中有用的新知识，即主观知识。主观知识再客观化于某一物质载体上，就成为文献，从而进入社会的公共知识、客观知识体系中，使原有文献中的客观知识得到继承和发展，实现了知识的进化（马恒通，2004）。

知识基因、知识DNA、知识细胞和知识体系是知识遗传与变异的结构要素（许志强，1990）。变异的知识再客观到某种载体上就进入社会，再经真实性和有用性的检验取得社会承认后，以形成的新的知识形态进入社会知识体系中，从而使社会客观知识得到进化。总之，人的知识结构对外来的文献知识信息经过吸附、同化、选择、建构，形成知识基因、知识DNA、知识细胞和知识体系，并实现客观化的过程，就是知识的遗传、继承与变异发展运动。

2）知识的流动实质上是一些思想在不同的知识主体之间的运动，科学技术知识流动的成因是，知识的进化和知识势差、区位势差。知识的进化是一个自然选择的过程，自然选择的结果就是优胜劣汰。只有那些优秀的知识才能流传和保留下来。也就是优秀的知识有自身的优越性，即比其他知识有优势，就是知识优势。知识优势是在知识流动过程中一条知识链相对于另一条知识链所表现出来的优势。知识优势包括知识存量优势和知识流量优势。知识优势来源于知识链在成员已有知识基础上的知识流动过程中的知识共享和知识创造（张睿鹏，2005）。

科学技术知识流动会形成聚集和扩散效应。科学技术知识流动过程中形成的聚集现象可以用“马太效应”来解释。聚集就是形成相对固定的知识群，从而形成相对固定的学术共同体，以及一些相对固定的研究者，成为相对固定的学术流派，形成学术研究的范式，有固定的学术领