项目编号：2018028

CALIS 全国农学文献信息中心研究项目
结题报告

项目名称：涉农学科引文网络的建立与应用

项目关键词：引文网络，耦合网络，文献推荐，投稿推荐，主路径分析

项目单位（盖章）：吉林大学图书馆

通信地址：长春市西安大路 5333 号吉林大学农学图书馆 130062

(详细地址含邮编)

项目主持人：谭智敏

联系电话：15843149319

电子邮件：tanzm@jlu.edu.cn

提交日期：2019.05.06

项目结题验收单

1 专家验收表（主持人所在单位组织 3-5 名专家对项目进行验收、自评。）

项目名称	涉农学科引文网络的建立、分析与应用				
主持人	谭智敏	职务/职称	副研究馆员		
所在单位	（加盖公章）吉林大学图书馆				
专家意见	该项目以吉林大学涉农专业和国内九所农业院校在 Web of Science（WoS）核心合集发表的论文为研究对象，建立了相应的学科引文网络，并分析了引文网络特征，可应用于文献推荐和投稿推荐中，选题新颖，具有很好的理论和实践意义。 该项目组研究认真，研究方法合理。经评议，专家组一致同意该项目通过验收。 （如需要可增加页数）				
专家签字	许[illegible]芳	陈[illegible]宇	崔[illegible]娃		
职务/职称	研究馆员	副研究馆员	副研究馆员		

题目：涉农学科引文网络的建立、分析与应用

关键词:引文网络，耦合网络，文献推荐，投稿推荐，主路径分析

1 研究背景、目的及意义

（一）研究背景

随着“双一流”建设总体方案的提出及各种大学排行榜的发布，科研评价和学科服务越来越受到国家、地区和高校的重视[1-4]。无论是科研评价，还是科研服务中的投稿推荐或者文献推荐，都是以引文网络数据分析为基础的。引文网络是由文献间引用和被引用的关系构成的集合，描述了科学领域的发展、学科内的脉络关系，包含了由研究作者、关键词、期刊、以及基金等组成的研究群体，这些群体内具有相近或者相关的研究内容，并代表着某个学科的知识结构、分布状况及未来的发展趋势。由于时间的推进和知识的积累，引文网络变得越来越复杂，不能很清晰的表示出学科演化情况和相关科研工作者在学科演化过程中起到的作用。而文献计量学是图书情报领域的重要研究方法，用以定量处理海量引文数据，分析引文特征，给出客观的“量”的评价结果。所以需要运用文献计量学的方法对引文网络进行分析和处理，其关键是用定量化方法对文献特征进行分析处理，以协助研究人员快速把握学科的发展脉络，激发科研灵感，开拓研究思路，发现新的研究方法等，更好地做好图书馆的学科服务。

(二)研究目的及意义

在引文网络中，最基本的表达是以文章作为网络节点，这样的网络是以文章的发表为网络新增的节点，文章的引用作为引文网络的连接，这样就使引文网络逐渐扩增。这就使当前大多算法都集中到对引文网络的结构进行研究，而这些算法中更基础的数据就是引文网络节点的入度和出度，目前制约这些算法的主要问题是网络节点数据量太大，需要一个有效的方式简化计算量。通过网络节点入度和出度分布的研究，可以给出入度和出度主要集中在那些部分，可以直接用一个有效概率截断

网络节点数据，大大减少了计算量，因此有利于简化当前学科服务优化算法。对于科研服务，需要分析引文网络主要连接方式，也可以说是主路径分析。节点评价，既可以是对文章评价也可以是对单位评价，这样根据评价高低，就可以进行文献推荐，或者对单位的科研质量进行测评和预测。对引文网络入度的统计分布，也可以用来分析引文入度的变化，入度代表各个单位发表的文章的分布，通过对这个分析，可以给出较好的投稿推荐。对其演化规律研究，就可以预测某个单位今后发表文章的质量变化。对于这些科研服务，当前研究主要集中在应用算法上，比如 Aprior，PageRank 等，但大多数是由于运算量过大，无法进行更精确的计算和预测，这对应用有了很大的制约。对引文网络的入度和出度的分布进行系统的理论和实例研究有利于简化这些学科服务算法。对网络节点分布研究有利于加速当前科研服务算法的在线应用，具有重要的应用前景。

2 研究内容及方法（思路、方法、具体内容）

2.1 研究方法

通过下载 web of science 上发表文章数据，对下载的数据按照数据字段汇总整理、清洗和分析，建立相应学科引文网络，分析引文网络特征。编写程序按照研究内容统计所需数据。对所研究对象进行建模分析。通过马尔可夫模型建模分析，对学科之间的差异进行统计研究，并给出实例验证。通过统计引文网络的入度和出度分布，发现入度和出度分布偏离幂率分布的现象，并根据内在作用机制建立模型，并解释分布。分别应用到投稿推荐和文献推荐中。

2.2 主要思路

2.2.1 马尔科夫模型建立应用到科研评价指标修正

评价指标大多都是从科研产出结果来统计或者评价的，都没有考虑到科研基础对科研产出的影响。科研基础主要包括在科研上可以逐年积累的量，比如科研经费数量或者基金资助数量、科研团队内人员数量、科研仪器的积累量和科研经验的积累等。所有这些都会对后续的科研产出有影响。这时我们应该通过其相对发展速度

来看它的发展情况，也就是在做科研评价时，需要首先排除这些可以累积的科研基础因素再评价其相对发展，这样才能得到真正的发展趋势和真正的学科对比情况。

2.2.2 扩展指数模型对引文网络入度分布拟合分析

对某一个科研机构发表文章进行统计，可以给出科研机构当前的发表文章分布的一个判断，这个分布能对当前科研机构科研能力以及科研能力的变化给出一个比较合理的反映，对这个分布给出一个合理的分布公式，有利于对网络节点分布偏离幂律分布的机制进行分析。以国内九所农业类大学发表文章的统计数据为例，建立一个通用型的模型，希望可以给出当前偏离幂律分布现象的内部机制，以及模型对应用的影响。

2.2.3 扩散模型对引文网络出度分布拟合分析

引文网络节点出度对应着文章的引文，同时也是科研工作者关注的文章，通过对出度的分析可以了解科研工作者的关注方向，进而可以给与文献推荐。本文通过构建一个可调参数模型，对吉林大学农学部的历史数据进行分析，并用最近一年的引用数据进行验证。

2.3 主要研究内容

2.3.1 科研评价指标修正

马尔可夫链模型分析方法用于处理评价指标，可以消除这些由于学科基础不同引起的差异，并能根据其结果预测今后的科研产出情况。所以本文引进马尔可夫链理论模型用于科研评估，通过现有数据拟合马尔科夫模型中的转移矩阵，并计算其平稳分布，用来描述科研水平和学科发展的情况评估。对马尔科夫模型对吉林大学各个学科中基金论文和非基金论文[5-6]的被引频次进行分析。得到结果如图 1 所示，图中给出基金论文和非基金论文综合评价，从图中可以看出不同学科中基金论文和非基金论文的发展速度相差不多。这说明虽然基金论文的数量和被引频次占有绝对的优势，但发展速度和非基金论文相差不多，甚至有的学科比非基金论文的发展慢很多。

影响因子是一个国际上通用的期刊评价指标，所以能在高影响因子期刊上发表文章也能在一个侧面上反映文章的质量比较高。为了说明文章发表质量的变化，我

们统计了吉林大学发文总量前 10 的学科的发表文章的影响因子分布，结果如图 2 所示，横坐标代表影响因子的区间，纵坐标代表各年发表文章的数量，图中的不同颜色代表不同年份发表的文章。从图中可以看出各影响因子分布区间大体上按年份增加，文章数量也随之增加，这和总体论文数量增加是一致的，说明了吉林大学发表的论文数量增加的比较均匀，不同影响因子区间都增加。但相对增加的量就比较难直接看出，为了得到发表文章的质量增速，我们对影响因子做了一个平均，这样就可以用一个相对值来描述文章质量。同时用马尔可夫模型加以分析。马尔可夫模型的权重选择为各个区间的平均值，高于 15 区间权重选择为 20。马尔可夫模型分析结果如图 3 所示，不同学科的综合评定值基本持平，这说明，每个学科对发表文章的影响因子还是比较重视，并都得到了很好的提高。

本项目通过马尔可夫模型分析，构建消除科研基础差异的综合评价指标，以便更客观的评价基础不同的学科的科研影响力，同时马尔可夫模型的转移速率代表了学科中不同层次的团队发展速率的快慢，也可用其对学科的发展速度及未来的发展趋势进行预测。

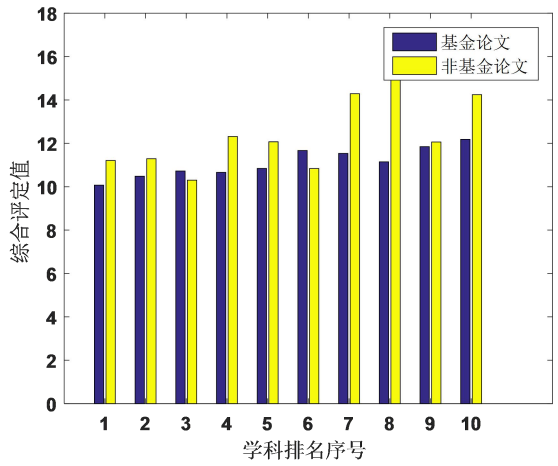

图 1. 基金论文和非基金论文的被引频次的马尔可夫评价结果

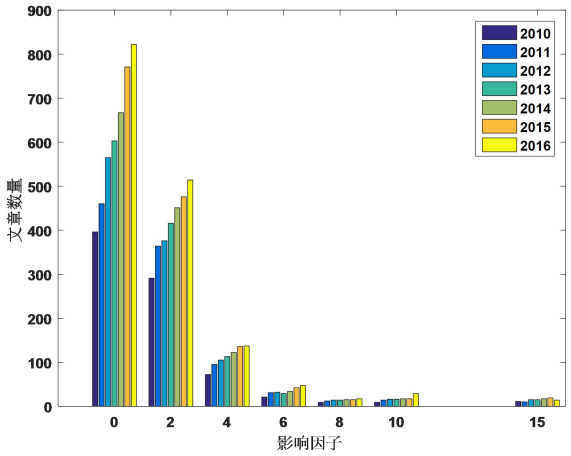

图 2.各年发表文章的影响因子分布

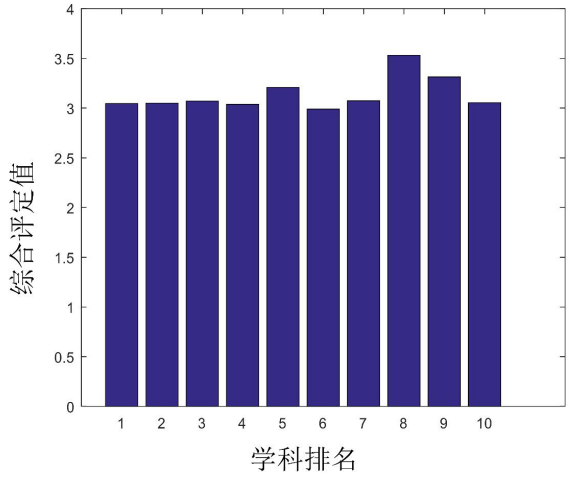

图 3.不同学科发表文章影响因子的发展速度

2.3.2 引文网络入度分布拟合分析

对 Web of Science 下载的发表文章数据按发表期刊的不同分别做统计分布，并按照对应期刊上文章刊载的数量做降序排列，并做归一化处理得到期刊发表的概率分布，可以得到如图 4（a）中蓝色圆圈描述的数据。随着文献的保存和传播的途径从纸质版过度到电子版，进而到目前有强大的搜索引擎和丰富的开放获取数据库，发表文章的统计分布也从满足布拉德福德定律[7,8]到满足幂律分布，近期的研究结果又出现一些偏离幂律分布的情况[9-14]。目前文献计量学内的解释主要归结为幂律分布。如果数据分布满足幂律分布，那么对概率和期刊序号分别做对数，结果应该是一条直线，对图 4(a)中的数据取对数，结果如图 4(b)中蓝色圆圈，红线是根据幂律分布拟合的结果，数据结果表明，在排名靠后的部分是偏离直线的。在正常

坐标下数据差别较小，但在取完对对数后，就可以明显看出其偏离幂律分布。已有研究者开始关注偏离幂律分布的现象，然而到目前为止还没找到一个合适的模型能精确拟合和解释偏离幂律分布的现象。为了更好地解释数据，本文通过抽象论文发表状态和状态转移速率，建立了一个类似动力学过的模型，并得到了扩展指数模型和指数模型求和的拟合方程，可以很好的拟合效果。

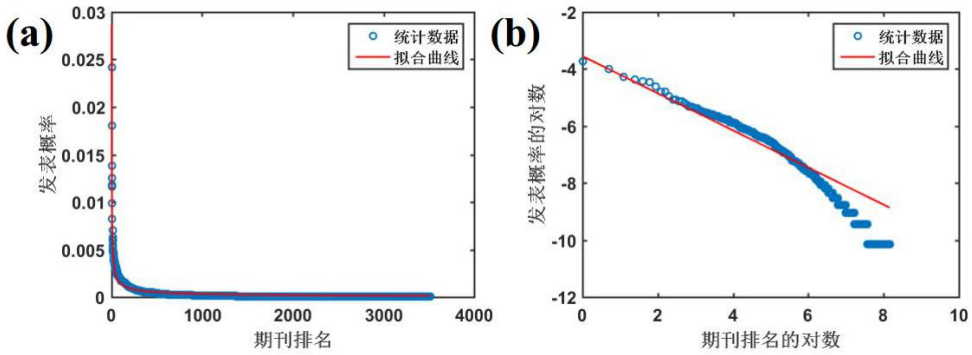

图 4. 期刊发表概率的统计分布图

蓝圈是数据统计结果，红线是拟合结果，（a）是正常坐标下的统计分布，（b）是双对数坐标下得到的结果。为了验证模型的普适性，我们对国内 9 所涉农学科的发表文章进行统计分析，结果如图 5 所示。从图中可以看出，对各个学校的统计结果都拟合的比较好，说明本模型的普适性。拟合结果如表 1 所示。从表 1 的数据可以看出，不同学校拟合参数变化较小，由此可以说明用这样的拟合参数组可以表征单个学科的发展状况。

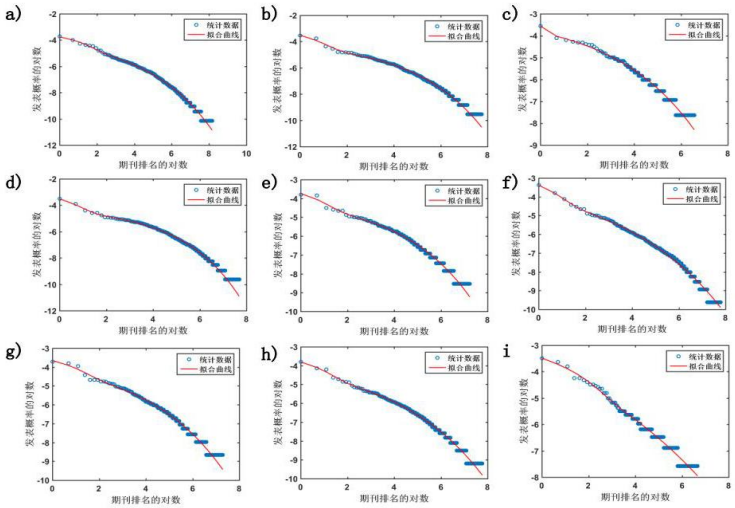

图 5. 九所学校发表期刊的统计结果及其拟合曲线

图中蓝色点是统计结果，红线是拟合曲线。（图中 a, b, c……，的顺序和表 1

中的顺序一致，学校的排序不分先后，按英文名称的字母排序）

表 1 各个学校数据通过扩展指数模型的拟合参数列表

学校名称	k_1	k_{-1}	β	A_1	A_2
中国农业大学	0.3102	0.3051	0.2819	0.0149	0.0253
华中农业大学	0.1055	0.6496	0.3416	0.0343	0.0189
吉林大学农学部	0.4877	0.8497	0.2629	0.0048	0.0581
南京农业大学	0.0510	0.7658	0.4042	0.0399	0.0151
东北农业大学	0.1122	0.5553	0.3296	0.0208	0.0193
西北农林科技大学	0.5579	0.6460	0.2648	0.0374	0.0337
山东农业大学	0.2628	0.4215	0.2982	0.0168	0.0292
华南农业大学	0.3658	0.4727	0.2673	0.0175	0.0245
云南农业大学	0.6027	0.1983	0.2532	0.0185	0.0349

通过建立的模型可以应用科研工作者的投稿推荐，通过建立的模型模拟，与完全随机模型对比，结果如图 6 所示，图中不同颜色代表初始投稿位置占总的杂志的比例，可见随着初始位置选取增加，投稿次数会明显的增加，但都比随机模型高很多，因此本模型有希望应用到个人投稿推荐系统，与其它算法联合可以得到更好的推荐准确率。

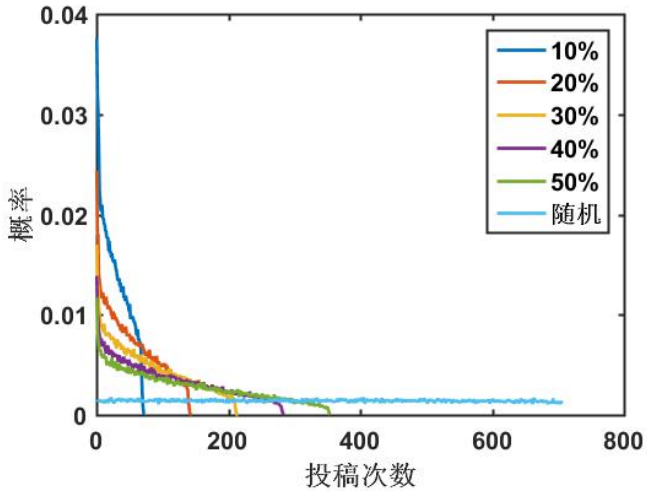

图 6. 概率随机模型用于投稿推荐与完全随机模型对比。

2.3.3 引文网络出度分布拟合分析

以 Web of Science 中吉林大学农学部发表文章中的引文数据为研究对象，通过类比自由扩散模型和引文发生过程，可以给出引文网络出度分布的拟合方程，方

程形式如下：

$$P(n) = A\frac{1}{1+\frac{n}{n_D}}\sqrt{1+\frac{n}{n_D}V^2}$$

式中的 A 是概率密度的归一化常数， V 对应观测体积，在文献引用模型中代表作者对某个具体问题检索时能精确到的范围。nD 对应扩散系数，代表用户在检索文献时在文献之间选择的能力。对吉林大学农学部引文数据的拟合结果如图 7 所示。结果显示，该模型可以很好的拟合统计分布数据。

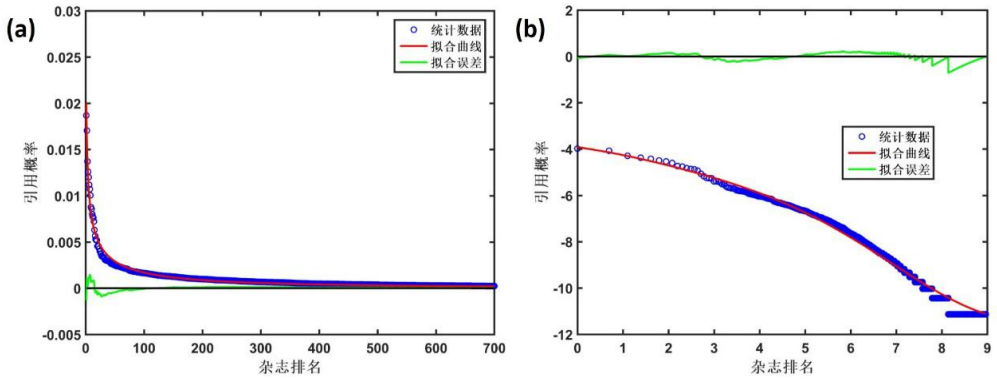

图 7. 引用期刊的统计分布以及拟合结果

蓝圈是数据统计结果，红线是本文建立模型的拟合结果，绿线是拟合误差（a）是正常坐标下的统计分布，（b）是双对数坐标下得到的结果。将本模型应用到文献推送，为了验证推送结果，本项目采用的对比模型是随机推送模型，模拟推送结果如图 8 所示。

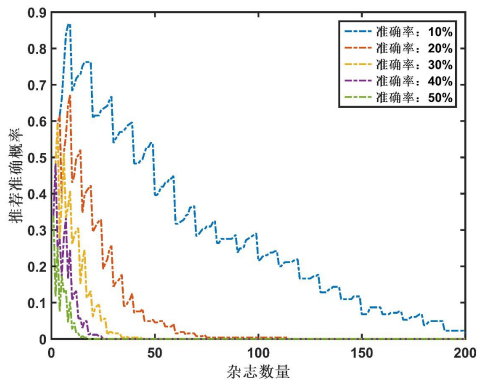

图 8. 推荐准确率结果

3 结论与建议

随着大数据时代的来临，科研评价、学科发展评价越来越受到大数据的影响，而引文网络数据是科研数据分析的主体数据，引文网络数据的研究有利于理解引文网络结构特征、科学计量指标的定义及修正，对促进文献计量学的发展也是非常重要的。对引文网络的定量分析也有助于了解学者的需求、做好数据库保障工作、更准确的文献推荐服务以及科研资源分配都有重要意义。

通过马尔科夫模型建立和分析，对常规的科研评价指标（论文数量、影响因子、基金论文等）进行修正，可以用于跨学科的评价。

通过对引文网络节点入度分布的研究，给出了入度分布偏离幂律的机制，并通过精确拟合的结果进行投稿推荐研究，与简单的随机推荐对比，大大的提高了推荐准确率。如果和其它推荐算法结合，有利于增加推荐算法的精度、减少推荐算法的计算量[15-16]。

通过对引文网络节点出度分布的研究，类比扩散过程，建立引文网络出度分布的拟合公式。并应用到文献推送，通过简单的概率模型推荐，就可以给出比完全随机推荐精度高出很多的推荐结果，如果配合其它推荐算法，有希望给出更精确的文献推送结果[17-18]。

4 项目成果（发表的文章、开发的软件、取得的实践效果等）

已发表论文：谭智敏,刘万国.基于马尔可夫模型的高校图书馆学科评价服务研究[J].现代情报,2019,39(03):150-156.

5 参考文献

[1]. 涂文菠. “双一流”政策下的国内高校图书馆学科服务探讨.[J] 大学图书情报学刊，2017，35（3）62-64

[2]. 刘雪立，魏雅慧，盛丽娜等. 期刊 PR8 指数：一个新的跨学科期刊评价指标及其实证研究.[J] 图书情报工作，2017，61，（11）116-123

[3]. 俞立平，张全，刘爱军. 不同学科多属性评价横向比较研究.[J] 图书情报工作 2014，58 (20) 100-105

[4]. 王雯霞，刘春丽. 不同学科间论文影响力评价指标模型的差异性研究.[J] 图书情报工作 2017，61（13）108-116

[5]. 李晓红，于善清，胡春霞，等. 科技期刊评价中应重视基金论文比的作用［J］. 科

技管理研究，2005，25(10) :138－139.

[6]. 许广奎，涂志芳. 两类学术评价指标比较研究. [J] 图书情报工作 2017，61（3）109-117

[7] Bradford B C. Source of Information on Specific subjects. [J]. Engineering, 1934(137):85-86.

[8] Price D J S. Networks of Scientific Pagers. [J]. Science, 1965, 149(368), 510-515.

[9]高晓培,袁军鹏,马峥,武夷山.科技期刊论文首次被引的幂律分布规律研究[J].情报学报,2015,34(07):693-700.

[10]俞立平,李磊.期刊影响力指标的幂律分布特征与差异研究[J].情报杂志,2015,34(03):85-88.

[11]毛国敏,蒋知瑞,任蕾,生冬梅,袁志祥,张放,宋胜合,葛之江.期刊论文被引频次的幂律分布研究[J].中国科技期刊研究,2014,25(02):293-298+307.

[12]彭旭辉.期刊引文幂律分布规律的实证分析——以经济学期刊为例[J].情报杂志,2016,35(07):185-189+157.

[13]耿志杰，王文鼐. 引文网络幂率分布特性的原因探析. [J]. 情报杂志，2009(11):15-17.

[14] 方爱丽，高齐圣，张嗣瀛. 引文网络的幂律分布检验研究. [J]. 统计与决策，2007(7):22—-24.

[15]. 王磊. 协同推荐技术及其在科技文献个性化推荐系统中的应用研究[D].南京理工大学,2007.

[16].张蓓. 基于组合策略推送算法的同城快递系统的设计与实现[D].北京交通大学,2016.

[17].周爱民.几种布拉德福分散曲线拟合模型的实证比较[J].情报杂志,2013,32(01):59-62.

[18].毛国敏,蒋知瑞,任蕾,生冬梅,袁志祥,张放,宋胜合,葛之江.期刊论文被引频次的幂律分布研究[J].中国科技期刊研究,2014,25(02):293-298+307.