



文献计量学理论与应用

李春英

北京大学医学图书馆

医学信息分析方法

- 定性
- 定量
- 定性与定量

内 容 提 要

- 文献计量学基础
- 文献计量学理论
- 文献计量学方法
- 文献计量指标体系设计
- 文献计量在科技评价中的应用

文献计量学基础——形成与发展

- 1911年，俄国学者瓦尔金，采用分析引文的方法，研究化学家在科学技术领域中做出的贡献
- 1917年，文献学家科尔和伊尔斯在《科学进展》，分析研究比较解剖学的文献，介绍了最基本的书目统计分析技术
- 1923年，休姆提出了统计书目学
- 1926年，美国统计学家洛特卡，研究著者与文献数量的关系
- 1934年，英国情报文献学家布拉德福，文献的集中与离散规律
- 1935年，美国语言学家齐普夫，研究了文献中词频分布规律
- 1958年，美国情报学家加菲尔德对引文分析进行了持续研究
- 1958年，贝尔纳、伯顿、开普勒提出了文献“半衰期”概念
- 1961年，美国科学家普赖斯提出了文献量的指数增长规律

文献计量学基础——概念

1969年由英国A. 普里查德提出文献计量学术语

- 文献计量学

是借助文献的各种特征的数量，采用数学与统计学方法来描述、评价和预测科学技术的现状与发展趋势的图书情报学分支学科。

文献特征

- 文献外部特征：著者、题目、刊名、书名、出版年、地址、出版类型、引用文献、文献序号等。
- 文献内部特征：分类号、主题词、关键词、代码等。

文献的外部特征

 [Heat shock response associated with hepatocarcinogenesis in a murine model of hereditary](#)

1. [tyrosinemia type I](#)

Angileri F, Morrow G, Roy V, Orejuela D, Tanguay RM.

Cancers (Basel). 2014 Apr 23;6(2):998-1019. doi: 10.3390/cancers6020998.

PMID: 24762634 [PubMed]

[Related citations](#)

缺氧诱导因子1 α 在肝癌细胞上皮-间充质化中的作用



原文索取



我的数据库

作者: 金炜东; 马丹丹; 蔡逊; 梅洪亮; 黄致远

作者单位: 广州军区武汉总医院普通外科, 湖北武汉 430070

出处: 中国普通外科杂志 2013; 22(7) : 885-889

相关链接: 主题相关 共引相关 作者相关

Accession Number: PREV200600549807

Document Type: Article

Title: Influence of fire and Juniper encroachment on birds in high-elevation sagebrush steppe

Author(s): [Noson, Anna C.](#) (anna.noson@umontana.edu); [Schmitz, Richard A.](#); [Miller, Richard F.](#)


Source: Western North American Naturalist 66 (3) : 343-353 JUL 2006


Language: English

文献的内部特征


缺氧诱导因子1 α 在肝癌细胞上皮-间充质化中的作用


Function of hypoxia inducible factor 1 α in epithelial-mesenchymal transition of liver cancer cells1


 原文索取

 保存到本地

 打印

 电子邮件

 保存到空间

 创建引文追踪器

流水号: 2013541922

作者: 金炜东; 马丹丹; 蔡逊; 梅洪亮; 黄致远

作者单位: 广州军区武汉总医院普通外科, 湖北武汉 430070

出处: 中国普通外科杂志 2013; 22(7): 885-889

ISSN: 1005-6947

国内代码: 43-1213/R

关键词: 癌, 肝细胞; 上皮-间充质化; 缺氧诱导因子1 α 亚基

摘要: 目的: 探讨缺氧诱导因子1 α (HIF-1 α) 在肝癌上皮-间充质化 (EMT) 中的作用。方法: 采用可调控HIF-1 α 表达的肝癌HepG2Tet-on-HIF-1 α 细胞系, 首先用real-time PCR与Western blot方法检测低氧环境中HepG2Tet-on-HIF-1 α 细胞EMT相关分子 (E-cadherin, vimentin, FSP-1) 及HIF-1 α 的mRNA和蛋白表达水平, 然后在常氧环境下, 采用强力霉素 (Dox) 诱导HepG2Tet-on-HIF-1 α 细胞HIF-1 α 过表达, 以及HepG2Tet-on-HIF-1 α 细胞经Dox处理后再转染HIF-1 α siRNA, 观察上述分子的表达情况。结果: 低氧处理后, HepG2Tet-on-HIF-1 α 细胞EMT相关分子及HIF-1 α 的mRNA和蛋白表达水平较常氧状态下均明显增加 (均 $P < 0.05$); 常氧环境下, Dox能诱导HepG2Tet-on-HIF-1 α 细胞HIF-1 α 过表达, 同时明显增加EMT相关分子的mRNA和蛋白表达水平 (均 $P < 0.05$), 但转染HIF-1 α siRNA后, Dox的诱导作用被取消。结论: HIF-1 α 促进HepG2细胞EMT, 并可能是肝癌基因治疗的有效靶点。

学科分类号: R341; R341.31; R394; *R735.7; R977.6

主题词[机]: *DNA结合蛋白质类; 蛋白质类/遗传学; 肝肿瘤/病理学; *核蛋白质类; *螺旋-环-螺旋构型; 氧; 中胚层; *转录因子

基金: 国家自然科学基金资助项目 (30371395)

文献计量学基础——发展

- 从上个世纪二十年代开始，把数学方法引入文献统计
- 从五十年代起将计算机技术用于文献引文分析使得文献计量学的理论和应用获得了巨大的发展

文献计量学与信息计量学、科学计量学的关系分析

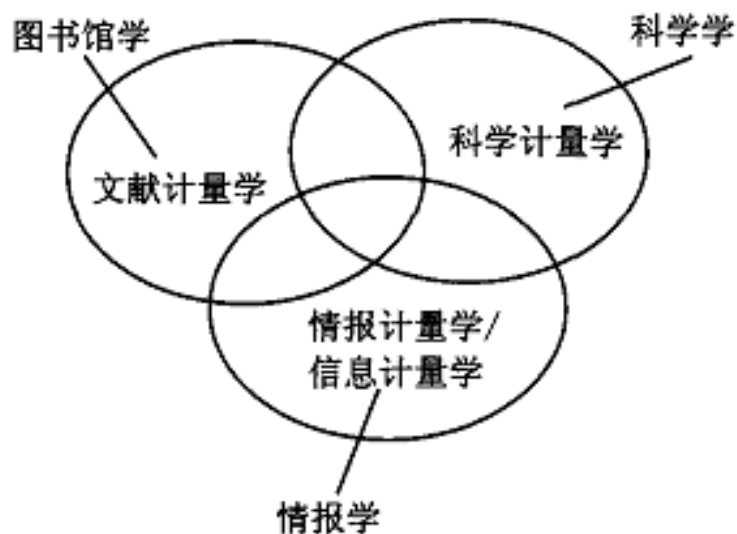


图 文献计量学、科学计量学和情报计量学（信息计量学）的联系与区别

- 依照它们的研究对象和目的来区分
- 计量学关注的三方面
 - 对象
 - 方法
 - 效果

文献、信息、知识、情报的关系

- 信息：客观世界中各种事物的变化和特征的反映以及经过传递后的再现。事物的状态、运动形式、运动规律以及事物与事物之间相互关系的表象。
- 知识：人类认识自然界、人类社会中各种现象、规律的信息反映，系统化、理论化的信息
- 情报：人们在一定时间内为一定目的而传递的有使用价值的知识或信息。
- 文献：人类的知识和信息，以文字、图形、代码、符号、声频、视频等形式记载到不同物质载体上的所有记录。

计量学理论的各门学科及其在科学评价中的应用

计量理论	相关学科	所属学科	研究内容	计量单元	应用
文献计量	文献计量学	图书馆学	科学研究活动产出分析	文献 (论文、著作、期刊、专利、数据库等)	科研能力、科研实力、科研竞争力、期刊、机构、人才、学科等评价
	信息计量学	情报学/信息管理学	信息源和信息传播过程	信息/情报	同上
	网络信息计量学	计算机科学/信息管理学/传播学	网络信息	网络信息和数据特征	网络评价与管理、网站评价与管理、数据库评价与开发管理等
科学计量	科学计量学	科学学/科技管理	科学研究活动的基本特征和规律	科研活动的基本特征和规律	学科评价、科学法制优先领域选择、科研资源配置
知识计量	知识计量学	知识经济学/知识管理学	知识单元层次考察知识的投入与产出、流量与存量、生产与应用、传播与分配等	知识体系	知识评价、科研成果学术价值评价、知识测度、知识生产与分配计量、知识投入与产出分析、知识资本与知识资源管理等
经济计量	经济计量学	经济学	从经济学角度考察科研投入(成本)、产出(效率和效益)等	科研成本效益分析	科研绩效评价、科研活动成本效益分析、科研资源配置

文献计量学基本理论

前苏联著名情报学家米哈依洛夫

(А. И. Мих айлов)指出：“当前，已发表文章的增长、老化和离散规律，理所当然地被视为标志科学文献发展的最根本的规律。”

文献计量学理论

- 文献增长规律
- 文献老化规律
- 洛特卡定律
- 布拉德福定律
- 齐普夫定律
- 文献引用规律

文献计量学理论——文献增长规律

- 科技文献指数增长规律

$$F(t) = ae^{bt}$$

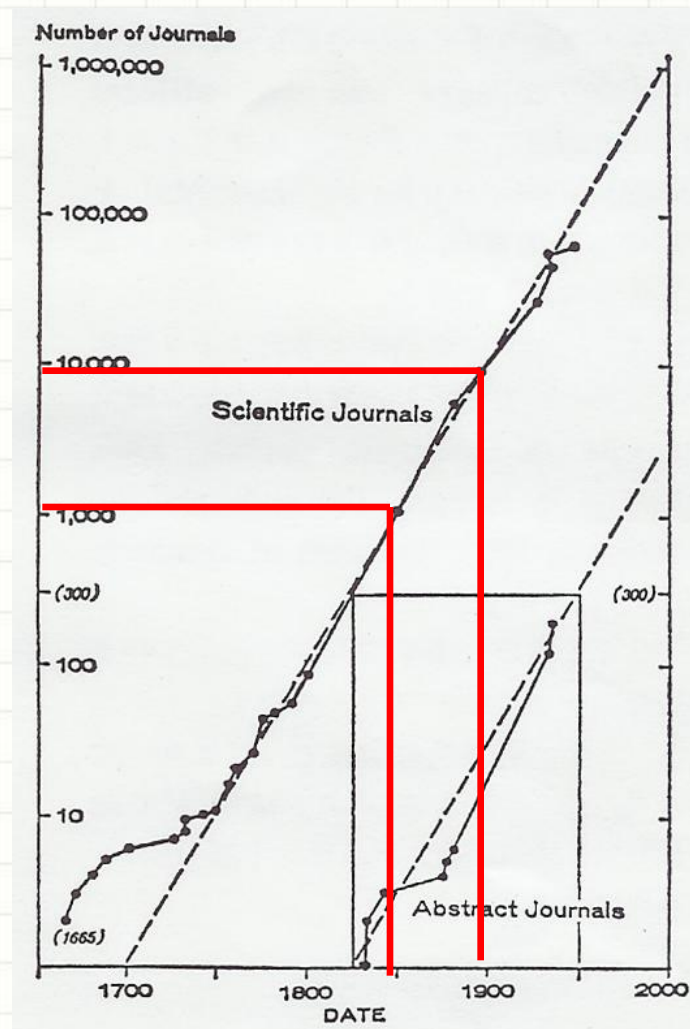
$F(t)$ 是时间 t 的函数，表示 t 时段的科技文献量；

a 为条件常数，表示统计的初始时段 ($t=0$) 时的科技文献量；

e 为自然对数的底数；

b 为时间常数，表示科技文献的年增长率 $r\%$ 。

研究表明，科技期刊数量和期刊文献数量均呈现出“按指数增长的规律”



科技文献指数增长律背后的故事

- ◆ 1949年，普赖斯在新加坡执教时，负责保管一整套《伦敦皇家学会哲学论坛》。由于十年一叠地放在床头书架上，使得杂志靠墙排成指数曲线状，这个现象被他意外地抓住了。
- ◆ 1950年，普赖斯回欧洲后向荷兰阿姆斯特丹的国际科学史大会提交了他的第一篇有关科技期刊按指数增长的文献计量学论文。该论文不仅标志他从数学和物理学转向了文献史研究，而且也成了他成长为科学计量学之父的起点。

指数增长模型分析-优点

正确性

- 正如普赖斯所指出“指数曲线的存在，显然具有普遍性和长期性。”因此，科学文献的指数增长定律具有较大程度的正确性，并获得了人们的公认。
- 从数学上分析和从统计实例来看，指数函数正是科学文献量随时间而增长的数学表示，符合一定历史年代的科学文献统计结果。

指数增长模型分析-缺点

局限性

- 并不是所有学科文献总是按指数规律增长
 - 科学文献并不总是按指数函数关系增长，它与所研究的文献的学科和时间有关。
- 不能预测未来
 - 按照指数函数的变化规律，随着时间的推移，科学文献的增量会趋向无穷大。显然，人类对科学研究的投入很难满足科学文献无限增长的要求。因此，这是不现实的。
- 科学文献指数增长规律相对于某年的文献累积量，并非指文献的增加量

文献逻辑增长模型

- 科技文献逻辑增长规律

$$F(t) = k / (1 + ae^{-bt})$$

$$(k > 0, a > 0, b > 0)$$

$F(t)$ 是时间 t 的函数，表示 t 时段的科技文献量；

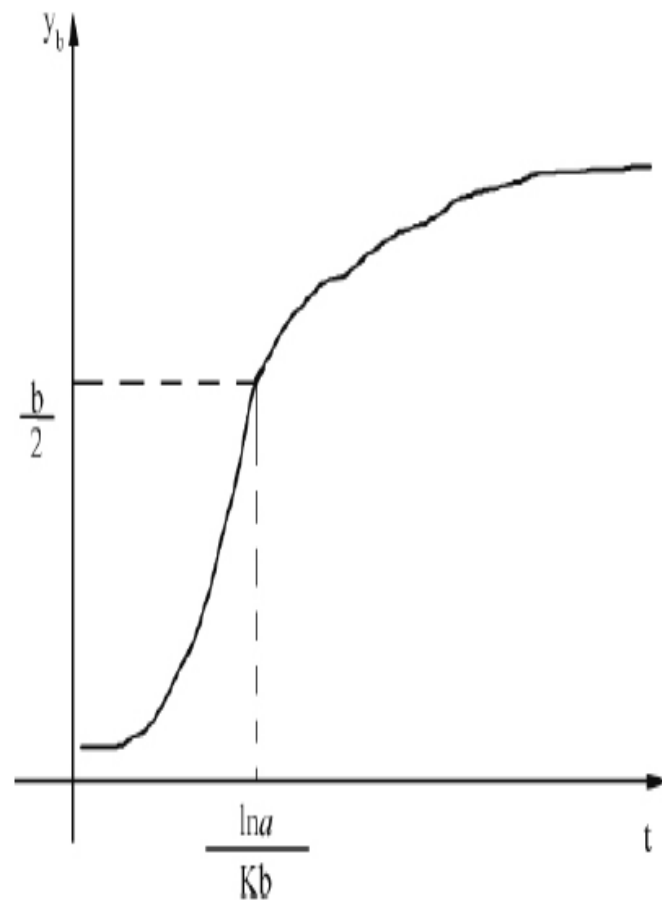
k 为当 t 趋向无穷大时的文献累积量；

a 为条件常数，表示统计的初始时段 ($t = 0$) 时的科技文献量；

e 为自然对数的底数；

b 为时间常数，表示科技文献的年增长率 $r\%$ 。

研究表明，文献的增长是分阶段的，每个阶段的增长模式并不相同：①指数增长；②增长率开始变小；③缓慢增加，趋近于一个极限值



文献计量学理论——文献老化规律

概述：

科技文献产生后，随着时间的推移，其流通及利用情况会发生变化：

- 有些文献的内容会被证明是不可靠的或错误的。
- 有些文献信息的内容尽管仍是正确的，被新的文献形式所替代，导致原有文献逐渐很少被人使用。

科技文献的这种逐渐失去使用价值而不再被人们利用或越来越少地被人利用的现象就是科技文献的老化现象。

科技文献老化规律

- 文献老化既是一种客观现象,又是一个复杂的动态过程
- 文献老化的原因: 科学知识不断增长和更新
- 科技文献老化研究实质: 对科学知识修正速度的探索

科技文献老化规律

- 科技文献老化的研究方法
 - 文献数据统计分析方法
 - 引文分析方法
 - 数学方法
 - 综合分析方法

科技文献老化规律度量指标

➤ 半衰期

- 所谓文献的“半衰期”，是指某学科(专业)现时尚在利用的全部文献中较新的一半是在多长一段时间内发表的。
- 1958-英国贝尔纳；1960-美国巴尔顿和凯普勒

➤ 普赖斯指数

- 在某一知识领域内，把对年限不超过5年的文献的引文数量与引文总量之比当作指数，用以量度文献的老化速度和程度
- 1971--普赖斯

➤ 剩余有益性指标

- 某一年份某一期刊被用户所利用的文献数
- 英国，B.C.布鲁克斯

科技文献老化规律

□ 科技文献老化的度量指标

■ 半衰期与普赖斯指数

- 一般来说，某一学科或领域文献的“普赖斯指数”越大，半衰期就越短
- 文献的“半衰期”只能笼统地衡量某一学科领域全部文献的老化情况
- “普赖斯指数”既可用于某一领域的全部文献，也可用于评价某种期刊、某一机构，甚至某一作者或某篇文章的老化特点

科技文献老化规律

□ 科技文献老化的数学模型

■ 巴尔顿——凯普勒老化方程

$$y = 1 - \left(\frac{a}{e^x} + \frac{b}{e^{2x}} \right)$$

□ $a + b = 1$

□ y 为经过一定时间该学科领域尚在利用的文献的相对数量

□ x 为时间，以10年为单位

科技文献老化规律

□ 科技文献老化的数学模型

- 巴尔顿——凯普勒老化方程
- 当取 $y=0.5$ 时，可以计算出文献的半衰期。巴尔顿等人据此测算出生物医学、冶金工程、物理学、化学工程、社会学等12个学科的文獻半衰期，其结果如下所示

学 科	半衰期	学 科	半衰期
生物医学	3. 0	生 理 学	7. 2
冶金工程	3. 9	化 学 学	8. 1
物 理 学	4. 6	植 物 学	10. 0
化学工程	4. 8	数 学 学	10. 5
社 会 学	5. 0	地 质 学	11. 8
机械工程	5. 2	地 理 学	16. 0

科技文献老化规律

□ 科技文献老化的数学模型

■ 巴尔顿——凯普勒老化方程

- 只考虑到“老化”而没有考虑到文献的增长，而文献的增长正是促成文献老化的重要因素
- 这个公式也较复杂，难以进行准确的实际计算

文献计量学理论——洛特卡定律

1926年，美国人口统计学家洛特卡发现了科学工作者与其论文数量分布规律

$$f_x = \frac{C}{x^2}$$

其中： x 为科学工作者所发表的论文数量； f_x 为发表了 x 篇论文的科学工作者占被统计该学科科学工作者总数的百分比； C 为常数，即为仅发表一篇论文的科学工作者所占的百分数。

$$f_1 : f_2 : f_3 : \cdots : f_n = 1 : \frac{1}{2^2} : \frac{1}{3^2} : \cdots : \frac{1}{n^2}$$

写了2 篇论文的科学工作者人数大约是写了1 篇论文科学工作者人数的1/4；
写了3 篇论文的科学工作者人数大约是写了1 篇论文科学工作者人数的1/9；
……写了 n 篇论文的科学工作者人数大约是写了1 篇论文科学工作者人数的
 $1/n^2$

洛特卡定律的限制

- 学科之间的性质差异，导致了不同学科作者著述行为的差异，从而对洛特卡定律在不同领域的应用产生影响
 - 物理学、化学、数学、经济学、人文科学等学科的 n 值接近2，比较符合洛特卡的平方反比定律
 - 生物学、计算机科学的著述分布不严格符合平方反比定律
 - 工程技术领域的 n 值则偏离较远，技术科学的 n 值会增大，即高产作者的比例降低，而规模较大、合作程度较高的学科 n 值会变小
- 该定律的适用有一个的前提，那就是对所研究的学科必须相对稳定，研究的论文时间区间必须足够长，研究的作者数目必须足够大，否则对该定律必须作相应的修正

文献计量学理论——布拉德福定律

- 文献集中和分散定律

- 布拉德福发现学科文献在期刊中的分布是有规律的，少量期刊集中了大量某学科文献，而其他期刊则很少出现该学科文献。他将期刊按发表学科论文的数量排序，划分出对学科最有贡献的核心区和随后的若干区。
- 后人证实了布氏揭示的学科文献在期刊中的分布规律是客观存在，将布氏发现的规律称为布拉德福文献集中与分散定律，将位于核心区的期刊称之为核心期刊。

产生背景

□ 文献的分散是普遍的客观现象

- 一个学科论文分散在其它学科的期刊杂志上屡见不鲜
 - 如何找出其分散的规律性是关键所在

□ 科学统一性原则

- 每一学科都或多或少与其它任何一个学科相关联
 - 对一个专家有用的论文，不仅会出现在这个专家所在学科的专业期刊，亦会出现在“其它学科”的期刊
 - “其它学科”期刊的数量，和其研究领域与“专家所在”学科的关系密切程度以及“专家所在”学科的论文在每种期刊中的登载率呈反比
- 按照某个学科论文期刊的载文率之高低来划分区域，每个区域中的期刊数量随着该区域期刊载文率的减小而增多

□ 文献统计研究是布氏定律产生的基础

文献计量学理论——布拉德福定律

- 如果将一定时间内(通常一年)的按某学科载文量等级排列的期刊划分为3个区,使每一个区所包含的相关论文数量相等,即恰好等于全部期刊发表的该学科文章总数的 $1/3$,则可发现:
 - 第一区涉及的论文来自数量不多但效率最高的期刊(称为核心区)
 - 第二区包含数量较大、效率中等的期刊(称相关区)
 - 第三区包含数量最大而效率最低的期刊(称外围区)
 - 核心区与相继各区的期刊数量成 $1:a:a^2$ 的关系

布拉德福定律应用的条件与局限

布氏定律只有充分满足以下列条件才能成立:

- 1) 论文的学科、专业领域或课题范围应当清楚地划定
- 2) 被分析的相关学科、领域或课题的期刊清单, 以及对这些期刊中刊载的相关论文的统计应当是充分的
- 3) 被分析的期刊的时间应当清楚限定, 以保证有关文献数据统计的一致性

具体来说, 布拉德福定律的局限在于

- 1) 一个很重要的无法模型化的参数: 学术制度变迁
- 2) 布拉德福定律的模型建立在了很多理想状态下

文献计量学理论——齐普夫定律

- 文献词频分布规律

- $f_r \times r = c$

- f_r ：頻次， r ：等級序號

齐普夫运用其“最省力法则”解释了这个定律。他认为，在任何语言中，凡是使用频率高的词，功能总是不会太大。因为词义本身在这个场合中价值小，因而传递它们所需要的“力”就不大。所以，词的出现频率与等级序号的乘积基本上稳定于一个常数。

齐普夫定律的适用性

- 一般来说，齐普夫定律较符合西文文献中词频分布的实际情况，定量揭示文献信息的词频分布规律
- 词频分布问题是很复杂的，因而使得上述公式的适用范围有一定的局限性
- 对出现频次特别高的词和特别低的词，不能圆满地反映其分布规律
 - 低频率的词，序号相同的很多
 - 高频率的词，序号相同的词随着频率的增高而越来越少

文献计量学理论——文献引用规律

- 在科学研究的过程中，必然要借鉴前人或他人的相关研究成果。科学文献体系中的每一份文献都不是孤立的，而是有着千丝万缕的联系，这种相互联系突出地表现为文献间的相互引用。

文献计量学理论——文献引用规律

- 文献引用关系

- 直接引用

- 间接引用

- 引文耦合：两篇或多篇文献同时引用一篇或多篇相同文章
 - 同被引：两篇或多篇文献共同被后来的一篇或多篇文献所引用
 - 自引：著者引用自己以前的著述

文献计量学理论——文献引用规律

- 研究内容

- 引文量的分析：文献类型、学科主题、语种、出版年代、引文来源等进行分析 and 描述

- 使用工具

- 《Web of Science》
- 《CSCD》

文献计量学方法

- 描述统计分析法
- 数学模型法
- 引文分析法
- 数据挖掘法
- 信息可视化法

文献计量学方法——描述统计

- 定义：是利用统计学方法对文献进行分析，以描述或揭示文献的数量特征和变化规律
- 类型：出版物统计、著者统计、词语统计、引文统计等

文献计量学方法——描述统计

- 步骤：
 - 统计调查：搜集获取研究对象的原始数据
 - 统计整理：数据计算、排序、图表显示等
 - 统计分析：统计数据的结论分析和误差分析
- 应用：单独运用或与其他分析方法结合，文献特征的描述，还用于科学学、预测学等

文献计量学方法——数学模型法

- 定义：是运用数学理论和方法，以数学表达的形式和符号来描述研究对象中的各种因素之间的数量关系，从而揭示其规律的一种研究方法。

文献计量学方法—数学模型法

□ 类型

— 研究对象的性质划分

- 必然现象模型：微分方程表达
- 随机现象模型：概率论和数理统计
- 模糊现象模型：模糊数学理论
- 突变现象模型：拓扑学奇点理论和结构稳定性理论

— 表达形式划分

- 解析式与图像模型：函数关系或图像描述系统(两个变量以下)
- 方程组模型：多变量情形
- 图表模型：系统的状态变化

— 描述方法划分

- 集合论模型
- 概率模型
- 代数模型

文献计量学方法——数学模型法

- 基本步骤

- 1.按具体项目的目的和要求来确定研究目标
- 2.根据某种理论建立模型，并找出其内部子系统和元件遵循的一般规律
- 3.根据一定时期的实际统计或样本资料，确定方程式的各个参数
- 4.进行模拟试验，检验模型的功能及可靠性
- 5.根据提出的数学模型进行预测和决策

文献计量学方法——引文分析法

- 引文分析定义：是利用各种数学及统计学的方法和比较、归纳、抽象、概括等逻辑方法，对科学期刊、论文、著者等各种分析对象的引证与被引证现象进行分析，以便揭示其数量特征和内在规律的一种文献计量分析方法。

文献计量学方法——引文分析法

- 类型

- 引文的出发点和内容

- 引文数量上分析：评价期刊和论文
 - 引文间的网状关系或链状关系：揭示学科的发展与联系
 - 引文反映出的主题相关性文献：揭示科学的结构、学科的相关程度和进行文献检索等

- 获取引文数据的方式

- 直接法：从来源期刊中统计原始论文所附的参考文献
 - 间接法：引文分析工具中获取

- 文献引证的相关程度

- 自引分析
 - 互引分析
 - 三引分析

文献计量学方法——引文分析法

- 基本步骤
 1. 选取统计对象
 2. 统计引文数据
 3. 引文分析
 4. 得出结论
- 注意：选取学科中有代表性、权威的若干期刊，确定一定时间范围内的相关论文

文献计量学方法——引文分析法

- 测度指标
 - 引文率
 - 期刊引文量
 - 期刊被引量
 - 自引率
 - 被自引率
 - 影响因子
 - 即年指数
 - 学科影响因子
 - 学科即年指数
 - 引证系数与被引证系数
 - 引文耦合
 - 同被引

文献计量学方法——引文分析法

- 作用

- 测定学科的影响和重要性
- 研究学科结构
- 研究学科情报源分布
- 确定核心期刊
- 研究科学交流和情报传递规律
- 研究文献老化和情报利用规律
- 研究情报用户的需求特点
- 科学水平和人才评价

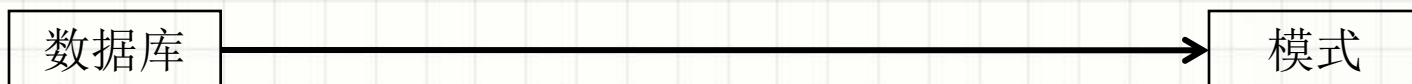
文献计量学方法——数据挖掘法

- 定义：从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。
- 对象：是超大型的数据库
 - 医院：HIS，电子病历
 - 商场：购物数据
 - 卫星天气资料

文献计量学方法—数据挖掘法

□挖掘的结果：是模式

- 从不断重复出现的事件中发现和抽象出的规律
- 对于集合的数据，可以用语言来描述其中数据的特性。如：
 - “如果成绩在81~90之间，则成绩为优良”
 - “如果成绩为81、82、83、84、85、86、87、88、89或90，则成绩为优良”。
 - 嗜烟者冠心病的发病率是不吸烟者的2~6倍。



文献计量学方法——数据挖掘法

- 数据挖掘分类
 - 数据库
 - 文本
 - Web信息
 - 空间数据
 - 图像和视频数据

文献计量学方法—文本挖掘法

- 定义：将文本型信息源作为分析对象，利用智能算法，并结合文字处理技术，分析大量的非结构化文本源，从中寻找信息的结构、模型、模式等各种隐含的知识。

文献计量学方法—数据挖掘法

- 数据挖掘算法—十大经典算法

C4.5	PageRank
k-Means	AdaBoost
SVM	kNN
Apriori	Naive Bayes
EM	CART

关联规则挖掘

- 关联规则挖掘：
 - 从事务数据库、关系数据库和其他信息存储中的大量数据的项集之间发现有趣的、频繁出现的模式、关联和相关性。
- 应用：
 - 购物篮分析
 - 捆绑销售和亏本销售分析。



“尿布与啤酒”

- 沃尔玛通过建立的数据仓库，定期统计产品的销售信息。
- 结果发现，每逢周末，位于某地区的沃尔玛连锁超市啤酒和尿布的销量很**大**
- 一些年轻的父亲下班后经常要到超市去买婴儿尿布，在购买婴儿尿布的年轻父亲们中，有30%~40%的人同时要买一些啤酒



之后该店打破常规，把啤酒和尿布的货架放在了一起。

Apriori 算法

基于两阶段频集思想的递推算法
找出频繁1-项集
找出频繁2-项集
用最小支持度、可信度等来衡量。

Tid	商品
1	面包, 牛奶
2	面包, 尿布, 啤酒, 鸡蛋
3	牛奶, 尿布, 啤酒, 可乐
4	面包, 牛奶, 尿布, 啤酒
5	面包, 牛奶, 尿布, 可乐

候选1-项集

面包 4
牛奶 4
尿布 4
啤酒 3
鸡蛋 1

频繁1-项集

面包 4
牛奶 4
尿布 4
啤酒 3

候选2-项集

面包 牛奶 3
面包 尿布 3
面包 啤酒 2
牛奶 尿布 3
牛奶 啤酒 2
尿布 啤酒 3

频繁2-项集

面包 牛奶 3
面包 尿布 3
牛奶 尿布 3
尿布 啤酒 3

频繁3-项集

面包 尿布 牛奶 2
面包 尿布 啤酒 2

应用实例：卓越购书



⊕放大
查看大图(放大)

摇摆:难以抗拒的非理性诱惑 (平装)

~ 奥瑞·布莱福曼 (作者), 罗姆·布莱福曼 (作者), 鲁刚伟 (译者), 何伟 (译者)

★★★★☆ (16 个用户评论)

市场价: ¥29.00

卓越价: ¥20.30 此商品可以享受[免费送货](#) [详情](#)

为您节省: ¥8.70 (7折)

VIP 价: ¥19.69 SVIP 价: ¥19.29

现在有货。

由卓越亚马逊直接销售和发货

配送至辽宁沈阳市和平区

如果您希望在10月29日收到商品,请在8小时6分钟内提交订单,并将送货方式选为[加急快递送货上门](#)。

若选择[快递送货上门-免配送费](#),您将于10月30日收到商品。 [了解更多](#)

商品促销和特殊优惠

- 每购买由卓越亚马逊提供的图书合格购物商品1件,另外购买1件[麦家最新力作《风语》](#) 可享受10%的优惠 [如何获得促销优惠](#)

[立即购买组合](#)

通常一起购买的商品

顾客购买此书时也通常购买[错觉:为什么我们视而不见、转身就忘或自命不凡?](#) - 约瑟夫·哈里南 平装 ¥19.60



+



卓越价合计: ¥39.90

[立即购买组合](#)

[查看发货和库存信息](#)

购买此商品的顾客也同时购买

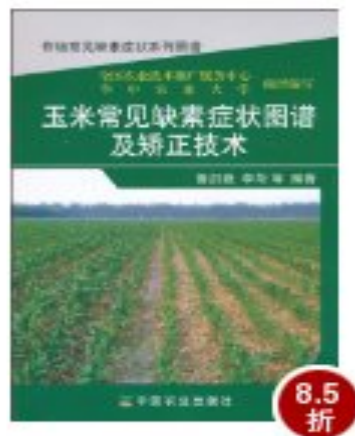


可笑的关联

joyo卓越
amazon.cn

尊敬的卓越亚马逊客户，

作为曾经购买或评价过 尹丽春 的 [科学学知识图谱](#)的人，您可能想知道现在可以订购 [玉米常见缺素症状图谱及矫正技术](#)了。 点击下面的链接预订，您只需支付 ¥ 10.20 (目录价的 {hash-get discount})。



[玉米常见缺素症状图谱及矫正技术](#)
鲁剑巍

市场价: ¥~~12.00~~

现价: ¥ 10.20

为您节省: ¥ 1.80 (15%)

由卓越亚马逊卖出并发货



添加到购物车

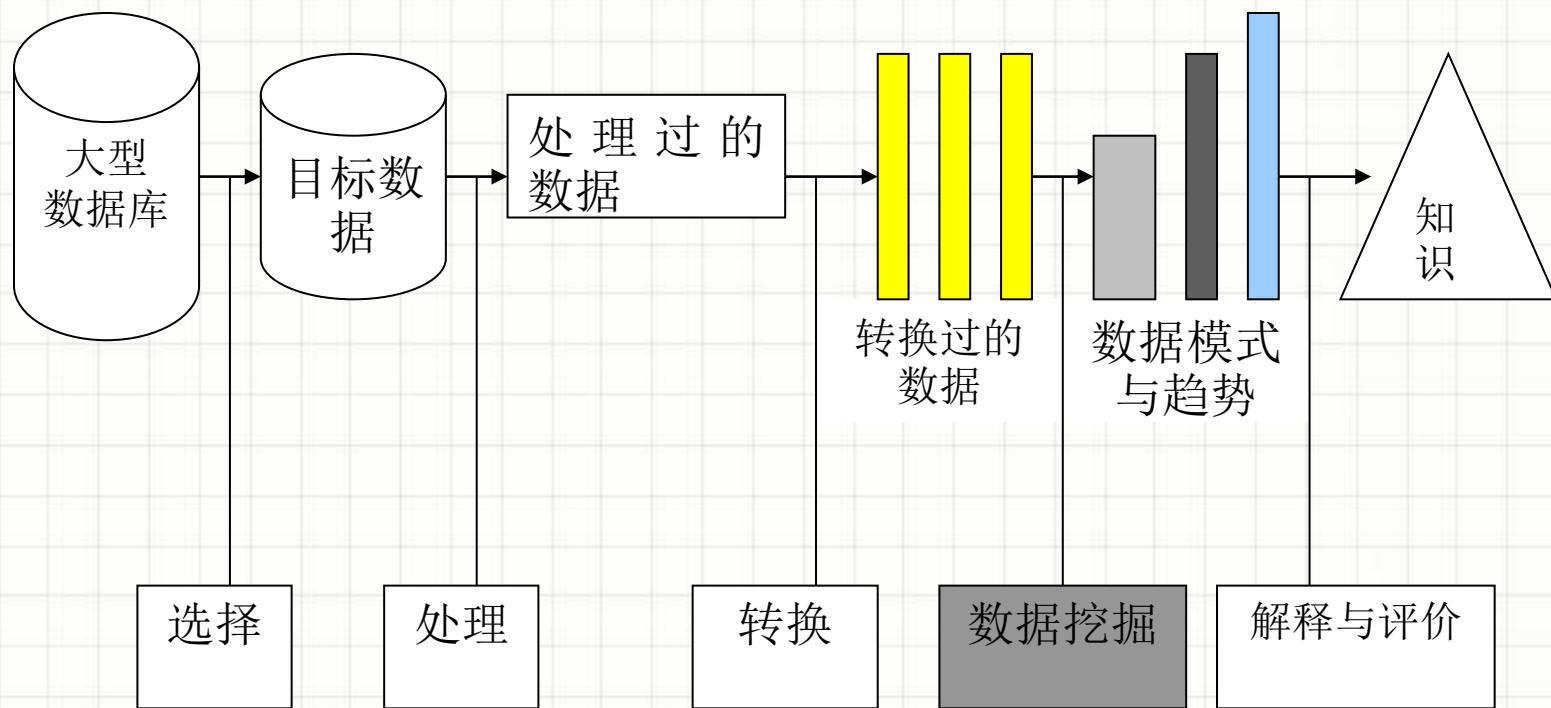
编辑推荐

《玉米常见缺素症状图谱及矫正技术》是作物常见缺素症状系列图谱。

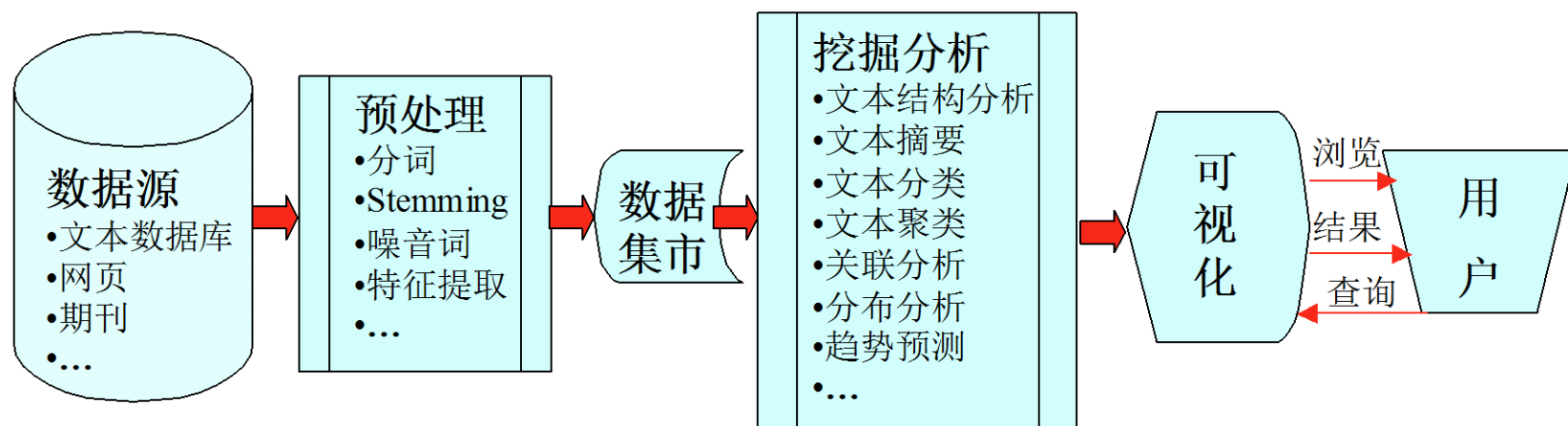
宿树兰 尚尔鑫 叶亮 段金廛.治疗痛经方药的关联规则分析.南京中医药大学学报, 2008, 24(6): 383-385

- 应用计算机检索中医方剂数据库（源于《中医方剂大辞典》）有关治疗痛经的217首方剂，以EXCEL 2000软件建立数据库，收录每首方剂中的单味药进行统计分析。
- 应用关联规则方法中Apriori算法分析方剂中药对的应用规律。
- 结果在治疗痛经的217首方剂中共使用427种药物2450频次。其中，使用频次在10次以上的依次为当归、川芎、延胡索、赤芍、香附等56味中药，使用总频次为1622次；
- 关联规则方法分析出使用频次在10次以上的药对当归-川芎、当归-白芍、当归-香附等389对。
- 结论运用用药频率统计与关联规则等数据挖掘方法，能较好地发现中医临床治疗痛经方药的用药规律，为临床遣方用药提供理论指导。

数据挖掘的过程



文本挖掘的一般过程



袁军鹏, 朱东华等. 文本挖掘技术研究进展[J]. 计算机应用研究(核心期刊). 2006, (02)

数据挖掘的过程

- 选择：根据某种标准选择或者切分数据。例如，将所有患有肺结核的病人的记录套录下来，形成该疾病患者的数据子集。
- 处理：包括清除和充实两个方面，由于数据是来自于日常工作中的记录，有许多冗余的和重复的内容，如病人的姓名可能在药局和实验室的数据库中都出现，有时还要从其他数据库中补充新的数据等等。
- 转换：删除那些丢失重要内容的记录，将数据分类（如按病人年龄分组），改变记录的格式（如将生日转换为实际年龄）等等。
- 数据挖掘：运用工具和算法，在数据中发现模式和规律（具体工具和算法后述）。
- 解释与评价：将发现的模式解释成为可以用于决策的知识，如预测、分类任务、总结数据库的内容或者解释观察到的现象。

数据挖掘的医学应用

- 药学

- 如利用趋势分析筛选药物，将某种药物在一定时期内的反应收集起来加以分析；
- 在大型化学数据库中自动寻找药效基团；
- 利用神经网络技术对世界卫生组织的药物副作用数据库的200万条报告进行数据挖掘，发现了药物间的相互作用。

- 病理

- 采用数据挖掘技术对显微标本中获得的大量数据(如计数、大小、形状特点、生理学评估、质地等数据)进行分析，总结出其中的关键性指标；
- 还可以对大分子及其化合物的电子显微镜三维致密重建图形进行数据采掘分析。

数据挖掘的医学应用

- 临床：HIS，EPR
- 通过临床日常工作和各项检查数据进行的数据采掘研究也逐年增加
- 应用大型的数据采掘软件,如DMSS软件(Data Mining Surveillance System)对ICU病房的微生物学数据进行分析,发现感染和抗药性**模式**上的变化
- 对医院感染和卫生检测数据进行数据采掘研究

文献计量学方法——信息可视化法

- 定义：利用计算机支撑的、交互的、对抽象数据的可视表示，增强人们对抽象信息的认识。其内涵是将数据通过图形化、地理化形象真是地表现出来，并且找到数据背后蕴含的信息

信息激增



文献计量学——信息可视化

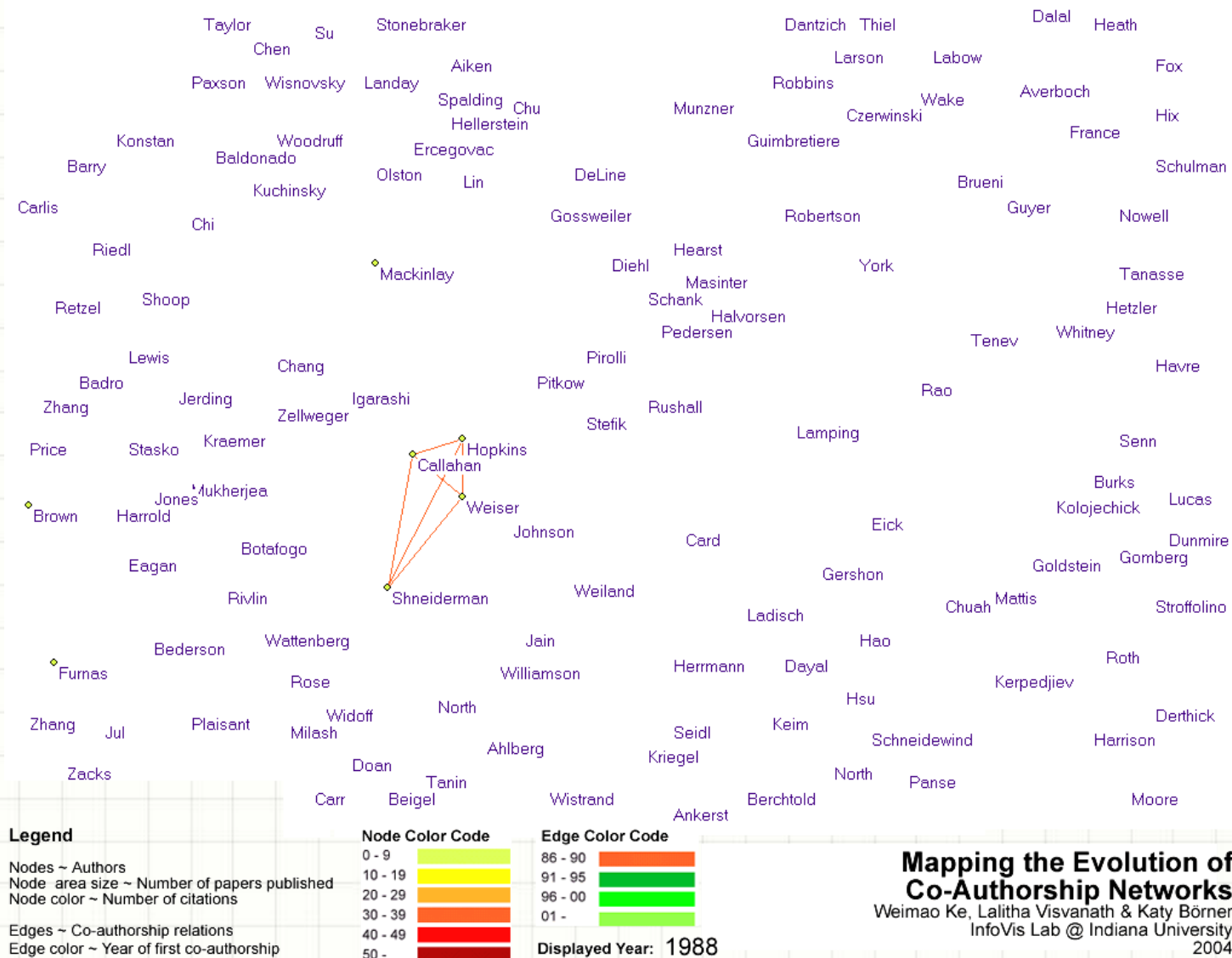
- 文献计量学、信息计量学/情报计量学——信息可视化
- 科学计量学——科学知识图谱

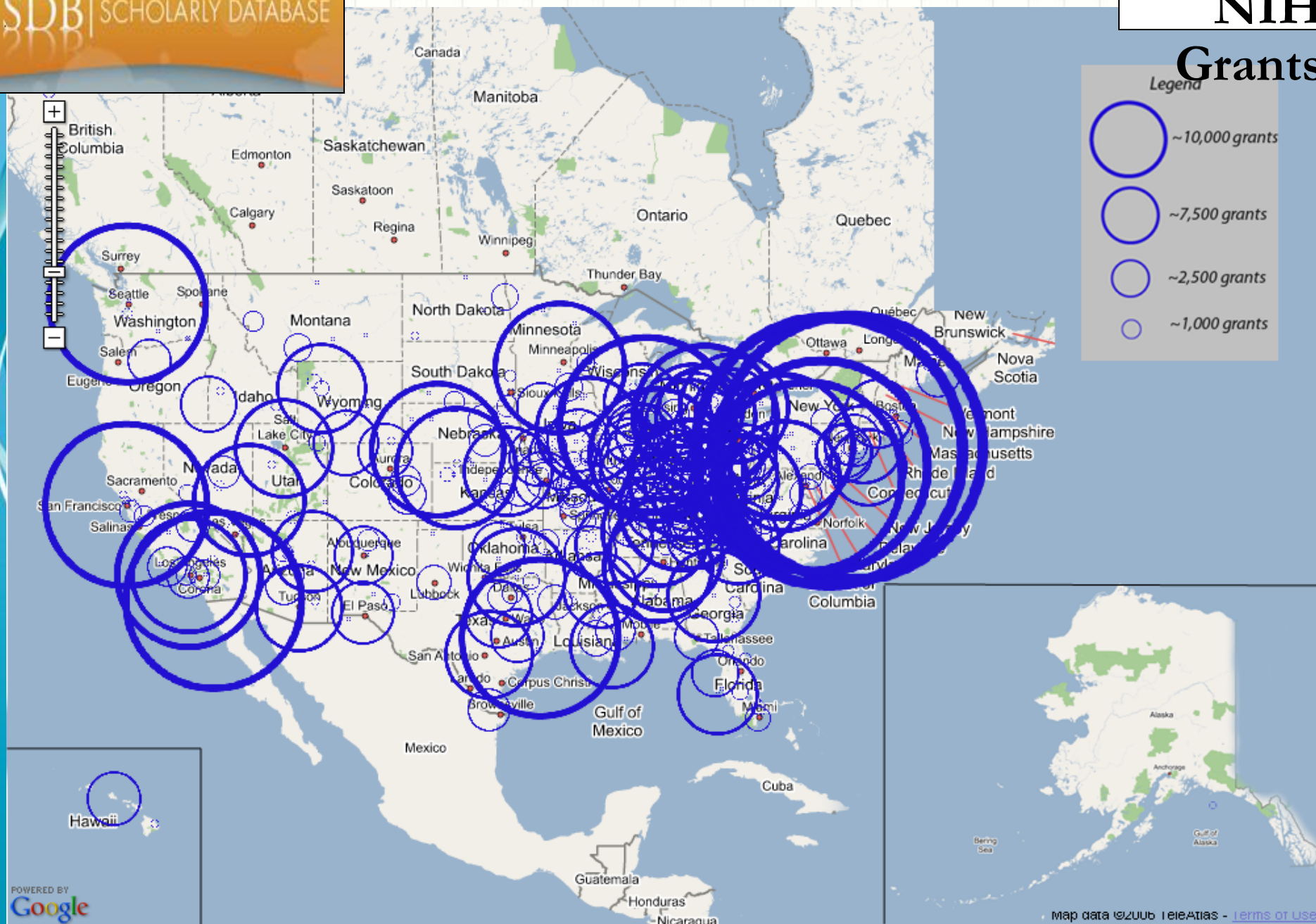
发展历程

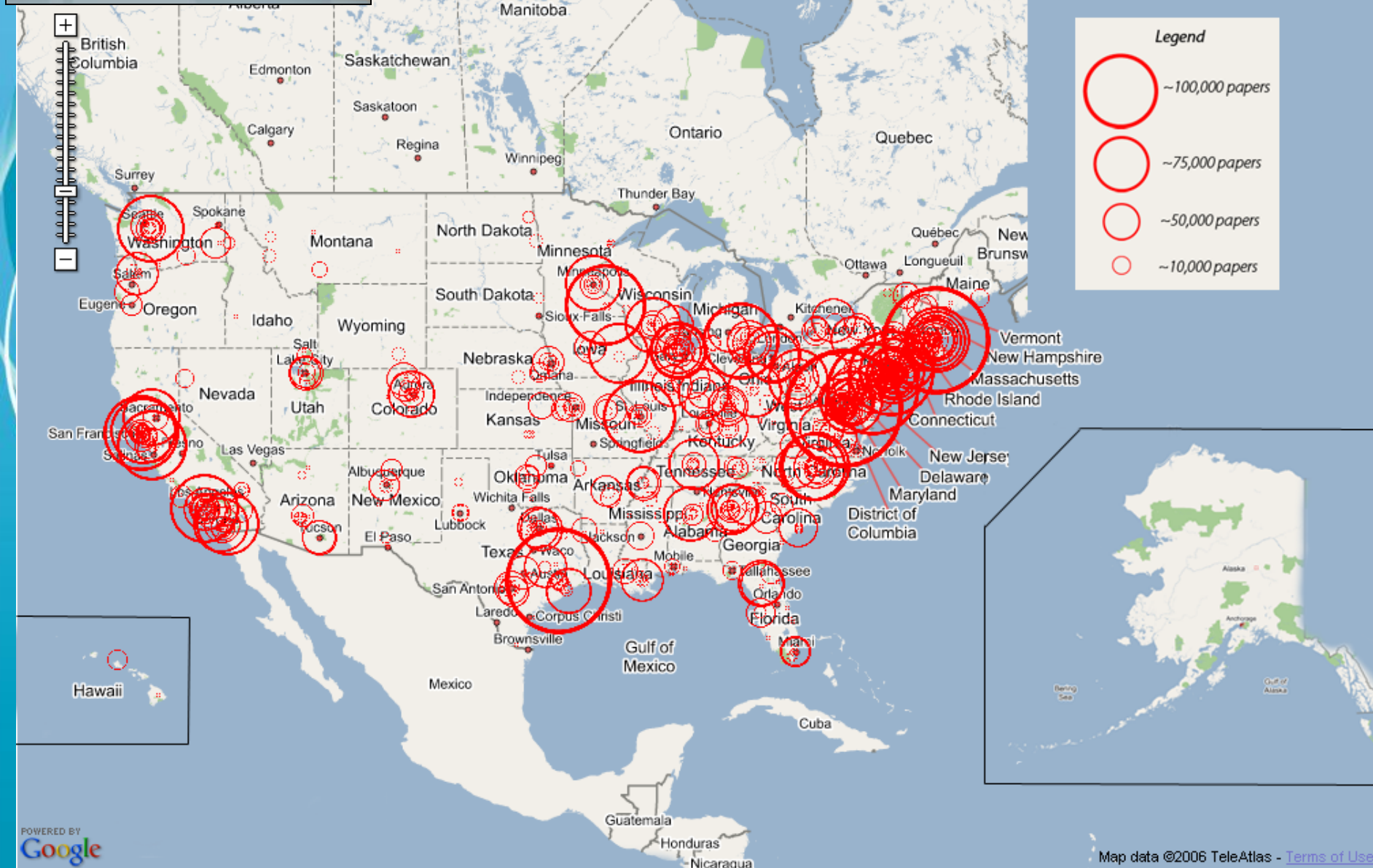
- 传统的科学计量学图谱以简单的二维图、三维图(如：线形图、扇形图、柱形图、散点图等)表达科学统计结果
- 1987年，美国基金委发展研究报告《科学计算中的可视化》开始长期资助科学可视化研究（三维立体可视化图谱）

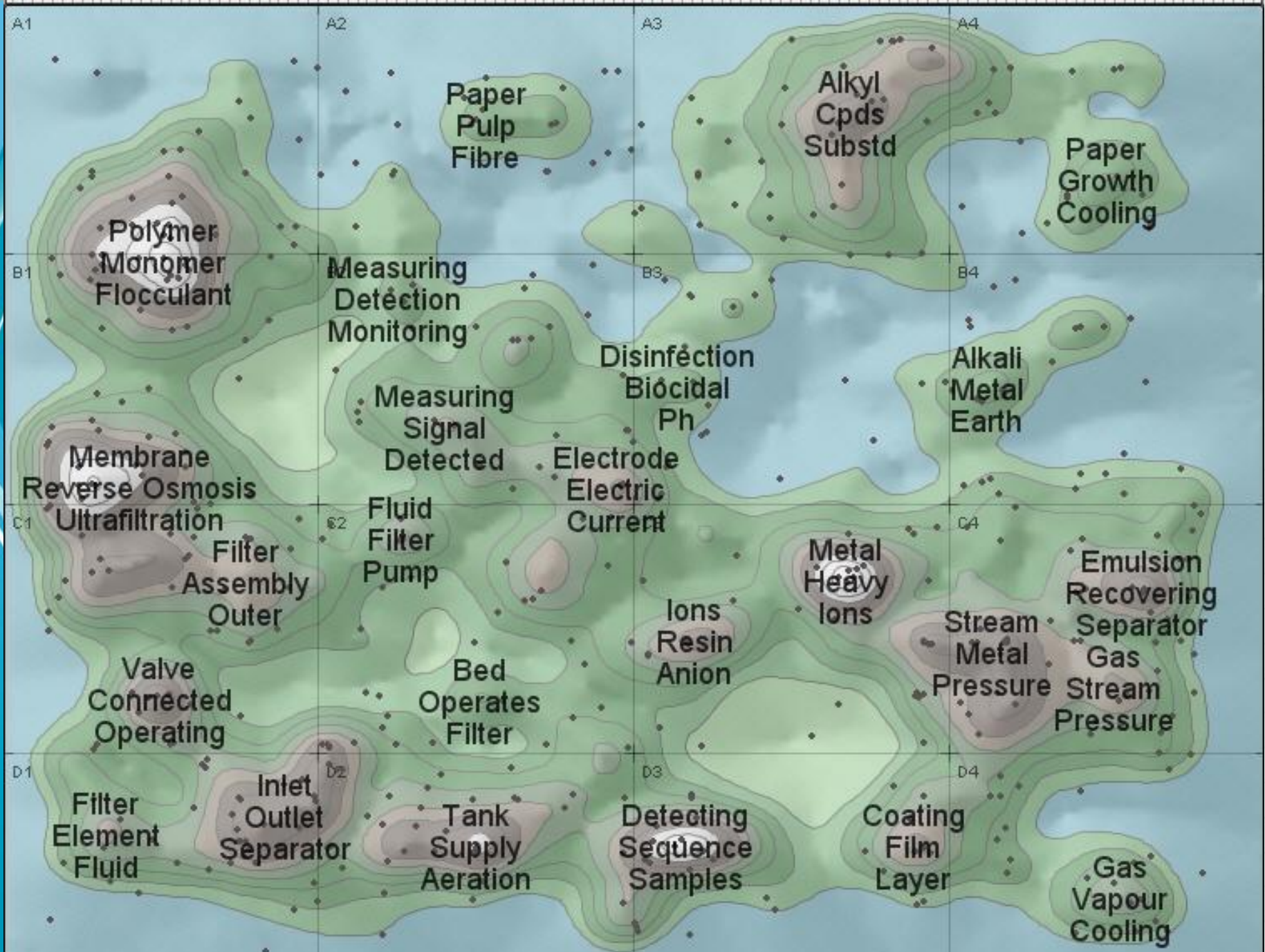
Mapping the Evolution of Co-Authorship Networks

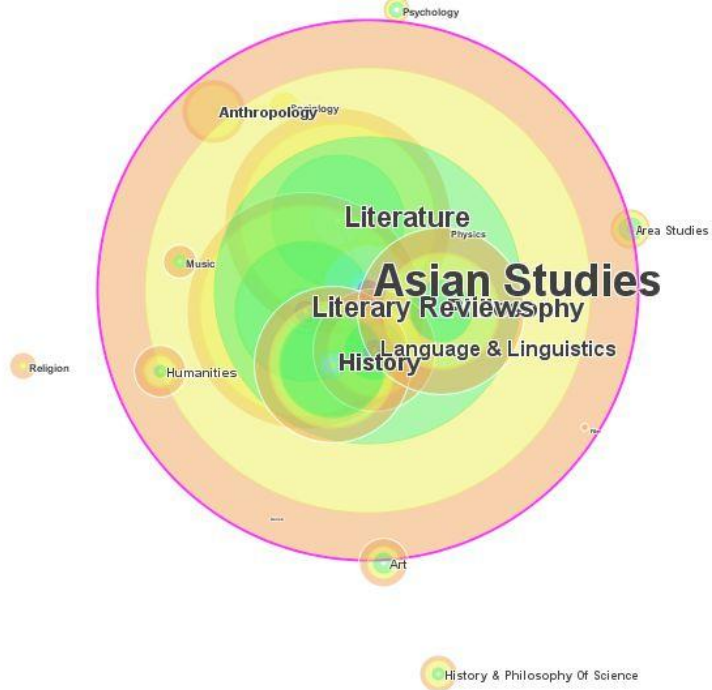
Ke, Visvanath & Börner, (2004) Won 1st price at the IEEE InfoVis Contest.



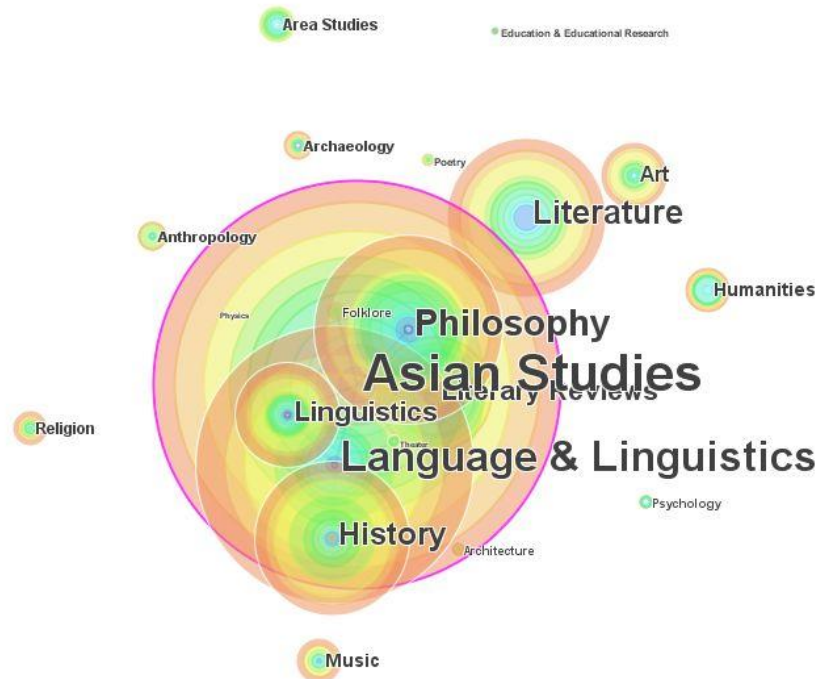








1984-1989



1990-1999

文献计量学方法——信息可视化法

- 过程

- 抽取：从文本信息对象中抽取外部特征和内容特征，建立专用数据表，准备原材料
- 转换：从文本对象中抽取特征是高信息维，要过渡到低可视维，先要进行必要的转换
- 映射：通过分析研究，要确定特定算法，启动程序可以将高信息维向低可视维自动映射
- 隐喻：合适的可视化模型可以隐喻文本对象的高信息维值，并发现规律的过程

文献计量学方法——信息可视化法

- 应用
 - 信息组织
 - 信息检索
 - 信息分析
 - 信息服务

文献计量学指标体系设计

□ 指标的分类

■ 总量指标

- 用来说明科技活动总体现象规模的统计指标
- 同类总体指标之和
- 决定所有科技调查资料的可靠性，因此对此类指标的命名必须标准、定义必须严谨、界定必须清晰、测量与计算必须规范，单位必须统一

■ 平均指标

- 按某个总量的平均标志来评价总体单位一般水平的统计指标

文献计量学指标体系设计

□ 指标的分类

■ 相对指标

- 以科技统计总量指标及相关的经济、社会统计的总量指标为基础进行符合数理统计理论的计算而形成的组合型指标
- 目的在于正确分析科技活动的结构特征、强度特征、变化趋势及科技活动现象与经济活动、社会活动现象之间的联系

文献计量学评价指标

文献计量指标体系设计

- 文献计量学指标

- 流通指标:

- 流通量一通常是指某一时间内文献借出的总数。
但从广义来说，凡是通过图书情报单位实现的信息交流，都叫“流通”，其计量就叫“流通量”。
目前仍以文献册数为其计量单位。

文献计量指标体系设计

- 流通指标

- 流通率——是一个相对指标，一般定义为：

- $$p = \frac{N}{M} \times 100\%$$
；其中N为统计时间内借阅的文献数量，M是投入流通的文献总数。

- 流通速度——是指单位时间内的流通数，其单位为册 / 时间。

- 这组指标的作用在于：能够反映流通工作量的大小、文献传递的速度和利用效率。

文献计量指标体系设计

- 藏书利用率

- 可表示为： $s = \text{全年出借册数} / \text{全馆藏书总册数} \times 100\%$ 。

- 图书周转率

- 是指一年中平均每本图书周转的次数，可表示为： $B = \text{全年图书借阅总册次数} / \text{全馆藏书册数} \times 100\%$ 。

文献计量指标体系设计

- 读者借阅率

- 在统计时间内平均每个读者所借的文献数，可表示为：

$$R = \frac{\text{借阅的文献总数}}{\text{借阅的读者人数}} \times 100\%$$

- 拒绝率：一般定义为：

$$H = \frac{\text{读者未借到的文献数}}{\text{读者要借的文献总数}} \times 100\%$$

文献计量指标体系设计

- 时差系数：定义为：

$$k = \frac{\text{当年文献的文摘条数}}{\text{该年文摘总条数}}$$

- 显然，k值越大则时差越小，说明其报道速度越快，可作为二次文献评价和选购的依据之一。

文献计量指标体系设计

- 情报吸收系数：一般定义为：

$$I = \frac{N}{M}$$

- M是统计时间内发表的文献总数，N是其中被利用的文献数。
- 该系数也可以相对于某一学科的文献或某种刊物而言。它是衡量情报被社会吸收利用的程度的指标。

科技期刊评价指标

文献计量指标体系设计

□ 科技期刊评价指标

- 科技期刊是反映科学技术产出水平的窗口，一个国家科技水平的高低可通过期刊的状况得以反映。
- 总被引用次数
 - 这是所评价期刊历年发表的论文在评价当年被其它期刊和该期刊本身引用的总次数，以表明该期刊在科学交流中被使用的程度。
- 影响因子（Impact Factor）
 - 可测度近年期刊的学术影响力。该项指标用论文平均被引率反映了期刊近期在科学发展和文献交流中所起的作用。
 - 定义：期刊前2年发表论文的被引次数占前2年论文总量的比例（相对数量指标）
 - 公式：影响因子= 某刊前两年发表的论文在该年的被引用次数 / 该刊前2年发表的论文的总数

文献计量指标体系设计

□ 科技期刊评价指标

■ 扩散因子

- 评估期刊真实影响力的学术指标，显示总被引频次所涵盖的期刊范围。

$$\text{扩散因子} = \frac{\text{总被引频次涉及的期刊数} \times 100}{\text{总被引频次}}$$

■ 平均引文率

- 可测度期刊的平均引文（引证文献）水平，考察期刊论文吸收他人学术思想的水平。平均引文率通常可以反映期刊吸收信息的能力以及科学交流程度的高低。
- 定义：在给定的时间内，期刊篇均参考文献量（相对数量指标）
- 公式：平均引文率=期刊参考文献总数/期刊论文总数

文献计量指标体系设计

□ 科技期刊评价指标

■ 即年指标

- 是表征期刊即时反应速率的指标，即该期刊在评价当年发表的论文，每篇被引用的平均次数。

■ 期刊载文量的地区分布数

- 这是衡量期刊论文覆盖和全国性的评价指标，我国按全国31个省（市）计，取近几年某期刊载文的地区分布数。

■ 期刊刊载的基金论文数

- 这是表明期刊所载论文学术水平和质量的一个重要指标，期刊载文的基金资助比例高，指示该刊学术水平较高。

文献计量指标体系设计

□ 科技期刊评价指标

■ 期刊被引半衰期 (cited half—life)

- 可测度期刊文献老化的速度。文献的半衰期受学科的内容、性质等因素的制约。一般来说，比较稳定的学科，其期刊半衰期要比正在经历重大变化的学科，或者说发展较快、较活跃的学科的期刊半衰期长；基础理论学科的期刊半衰期要比技术学科的期刊半衰期长；历史悠久的学科期刊半衰期要比新兴学科的期刊半衰期长。
- 定义：某期刊现尚在被引用的全部论文中较新的一半的年代跨度。

文献计量指标体系设计

□ 科技期刊评价指标

■ 期刊他引率 (non—self—cited rate)

- 可测度某期刊学术交流的广度、专业面的宽窄以及学科的交叉程度。
- 定义：期刊被他刊引用的次数占该刊总被引次数的比例（相对数量指标）。
- 公式：期刊他引率=被他刊引用的次数 / 被引用的总次数
- 这个指标是《中国科技期刊引证报告》最早提出来的，通常用于表征期刊科技交流中的范围和程度。

文献计量指标体系设计

□ 科技期刊评价指标

■ 期刊的国际化程度

- 根据海外作者来稿数统计。

■ 国际著者论文数

- 指期刊发表外国作者论文的数量。可测度期刊在国内外学术交流中的作用。
- 外国作者论文数包括外国作者独立撰写的论文和外国作者与国内作者合作的论文。

文献计量指标体系设计

□ 科技期刊评价指标

■ 期刊的国际化程度

- 根据海外作者来稿数统计。

■ 国际著者论文数

- 指期刊发表外国作者论文的数量。可测度期刊在国内外学术交流中的作用。
- 外国作者论文数包括外国作者独立撰写的论文和外国作者与国内作者合作的论文。

■ 平均作者数

- 来源期刊中每篇论文的平均作者数，衡量期刊科学生产能力的指标。

文献计量指标体系设计

□ 科技期刊评价指标

■ 学科扩散指标

- 指在统计源期刊范围内，引用该刊的期刊数量与其所在学科全部期刊数量之比。

$$\text{学科扩散指标} = \frac{\text{引用刊数}}{\text{所在学科期刊数}}$$

■ 学科影响指标

- 指期刊所在学科内，引用该刊的期刊数占全部期刊数量的比例。

$$\text{学科影响指标} = \frac{\text{所在学科内引用被评价期刊的数量}}{\text{所在学科期刊数}}$$

■ 文献选出率

- 按统计源期刊选取论文的原则选出的文献数与期刊的发表数之比。

文献计量指标体系设计

□ 标准规范化

- 执行国家有关标准和规范化程度（专家评定）。

□ 论文平均发表周期

- 指评价当年全刊所有论文投稿到发表所需时间的平均

□ 编校质量

- 指文字表达，标点符号的使用，图表编排和校对质量（专家评定）。

□ 装帧印制质量

- 指封面、版面设计与印刷、装订质量（专家评定）

□ 人才培养

- 期刊在发现和培养科研人才方面所起的直接或间接作用（报告叙述）。

文献计量学在核心期刊评价中的应用

- 某学科核心期刊是指刊载该学科学术论文较多的、论文被引用较多的、受读者重视的、能反映该学科当前研究状态的、最为活跃的那些期刊。

核心期刊概念的发展

- 二次文献(文摘、提录、索引)的核心期刊效应
 - 1967年，联合国教科文组织的一篇文章指出“从物理学和化学领域的重要文摘杂志中发现了一条规律，它们所列出的或编成文摘的75%的论文，仅来自它们所收摘的全部期刊的10%”

核心期刊概念的发展

● 流通量的核心期刊效应

□ 1969年，高夫曼（W. Goffman 美）、莫利斯通过统计分析，证实按期刊流通量数据的分布近似服从布拉德福文献分散规律，存在核心期刊效应。

核心期刊概念的发展

- 被引量的核心期刊效应

- 1971年，加菲尔德（Garfield, SCI创始人），在统计了2000种期刊中的1百万篇参考文献后发现，24%的被引频高的文章出自25种期刊，50%的出自152种期刊，75%出自767种期刊，而其余的被引文章则散布在数量大得多的期刊中。证明了被引文章在期刊上的分布也有一个较为集中的核心区与广为分散的相关区

核心期刊概念的发展

- 有许多研究表明，由上述因素派生的其他因素如：被引率、影响因子、即时被引率等也都具有核心效应。

核心期刊的作用

- 可作为图书馆期刊订购、导读和参考咨询、读者投稿，文献数据库选择来源刊，研究成果评价等方面的参考工具
- 核心期刊只是一种相对的统计的概念，核心期刊表只能起参考作用，不能起标准作用。
 - 从个体角度看，核心期刊上的文章未必每篇学术水平都高，非核心期刊上的文章未必每篇水平都低。
- 因此在评价研究成果时，还应根据单位或评价项目的具体情况，请学科专家来评审论文本身的学术价值。

核心期刊评价在中国

- 北京大学图书馆与北京高校期刊工作研究会的《中文核心期刊要目总览》
- 中国社会科学院文献信息中心的《中国人文社会科学核心期刊要览》及《中国人文社会科学引文数据库》
- 中国人文社会科学学报学会的《中国人文社科学报核心期刊》
- 中国科学院文献情报中心《中国科学引文数据库》及其来源期刊
- 中国科技信息研究所的《中国科技论文引文分析数据库》及“中国科技论文统计源期刊”

文献计量学在科技期刊评价中的应用

□ 科技期刊评价的模式

□ 客观评价

- 采用科学、公正、客观的评价指标体系对科技期刊的学术质量以及整体水平进行评估，这也是科学界共同提倡的、可以减少争议的公平做法

□ 主观评价

- 定量和定性相结合的方法，由专家进行最后的投票和打分

□ 合理的评价体系

- “以评促优，以评促改”

□ 科技期刊评价机构

- 从国家层面，统一组织，由权威的、中立的研究单位进行科技期刊评价和发布是比较可行的做法。

中国科技期刊评价指标

学术质量水平指标

国际竞争力水平指标

可持续发展潜力指标

影响因子

总被引频次

基金论文比

他引总引比

即年指标

进步指标

文章下载率

平均引文率

扩散因子

国际检索系统收录

国际优势学科

国际被引用次数

国际编委比

国外发行量

国际论文比

主编学术水平

主创人员开拓能力

办刊单位支持情况

期刊电子化水平

编辑出版质量

期刊声誉

参考文献:

- 1.李道苹.医学信息分析.北京:人民卫生出版社.2009
- 2.袁军鹏.科学计量学高级教程.北京: 科学技术文献出版社.2010
- 3.邱均平, 文庭孝.评价学: 理论、方法、实践.北京: 科学出版社.2010
- 4.中国医科大学教学课件
- 5.周宁丽.科学知识图谱-CiteSpace利用方法



谢谢!