中国科学院国家科学图书馆

青年人才领域前沿项目研究报告

项 目 名 称：科学前沿方法的研究及其规则的建立

项目负责人：李泽霞

项目完成人：李泽霞 杨帆 邢颖 黄龙光 李超

所 在 部 门：国家科学图书馆情报部

通 信 地 址：北京市海淀区中关村北四环西路 33#

电　　话：82626611-6161

Email: lizexia@mail.las.ac.cn

完 成 日 期：2010 年 7 月 12 日

中国科学院国家科学图书馆

2009 年 9 月制

目 录

一、研究背景及意义

目前的文献数据分析还主要基于统计和计量，而且在数据处理和分析的过程中，主要关注具有规模效应的数据簇，往往忽视甚至丢弃了文献数据中的孤立点数据，而这些数据有时却代表了某一种学科的突破性进展或新兴研究类型的出现，或是某些薄弱研究领域的异常警示。一些重要和敏感的此类信息有可能为科技决策者对科研的政策制定和学科的优先部署提供很好的参考价值。

目前情报研究中还没有很好的通过文献数据检测科学前沿的方法。通常采用关键词的词频和文章的共引等指标来度量和分析文献。同时这些具有规模效应的数据通常指示了一些较成熟、活跃的研究领域，通常反映了这些突破性研究进展已经到了研究的爆发发展阶段。研究过程往往忽略了新出现的和还没有形成聚合效应的数据，同时丢弃了很多有用的信息。

因此本研究拟尝试将孤立点挖掘来查找潜在的新兴研究方向和具有突破性的研究性进展，以此来挖掘文献数据中的潜在有用信息，并利用时序分析方法结合文献计量的相关指标建立判断有意义孤立点的相应规则。

二、项目研究过程及要点

2.1 科学前沿方法的调研

经过调研分析总结得出：应当通过三个方面对科学前沿信息进行全面监测，a）聚合特征明显的热点信息，目前大多数前沿信息都是以具有规模效应的数据簇来表现的；b）为了弥补聚合特征明显的数据对其他相对较弱的信息的掩盖，降低阈值来发现宏观结构下潜在热点信息，通过对微细结构的展现发现目前不是热点但即将成为热点的信息，对聚合特征相对较强的信息进行分析辨别以发现科学前沿；c）没有聚合特征的信息（孤立点或孤立点簇等），定义一定的规则来发现潜在的新兴研究方向 1。

[1]张英杰博士通过构造主题关键词包容矩阵，邻近矩阵和等价矩阵从优选聚类的角度筛出历年离散点，结合在国际空间站调研中的实践认识，对相关主题概念的演变进行了判读和研究，揭示了部分有意义的研究内容。

2.2 孤立点检测方法的分析和研究

2.2.1 孤立点的类型

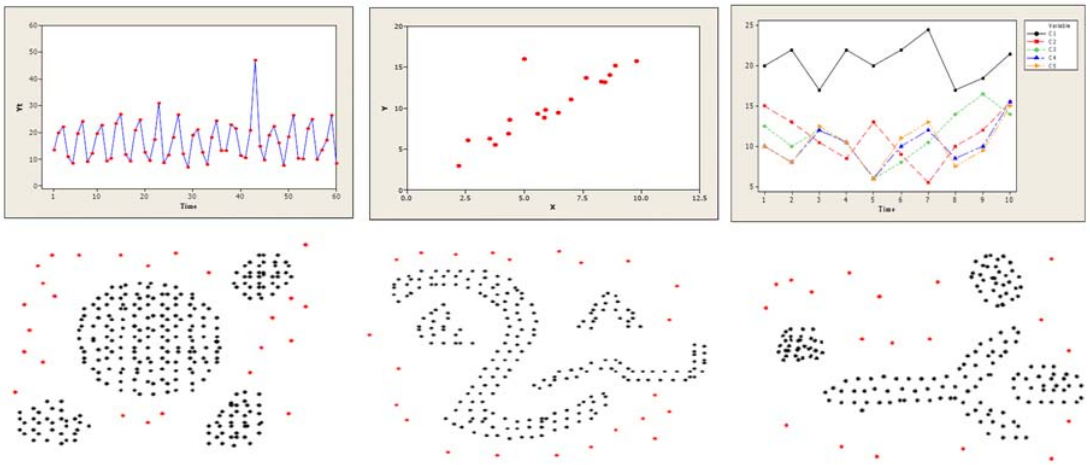

图 1 孤立点的类型

通常情况下，数据孤立值分为全局孤立值和局部孤立值两大类。全局孤立值是指对于数据集中所有点来讲，具有很高或很低的值的观测样点。局部孤立值值对于整个数据集来讲，观测样点的值处于正常范围，但与其相邻测量点比较，它又偏高或偏低。

孤立点的出现有可能就是真实异常值，也可能是由于不正确的测量或记录引起的。如果孤立值是真实异常值，这个点可能就是研究和理解这个现象的最重要的点。

2.2.2 孤立点检测的方法

- **基于统计的方法-**基于统计的孤立点检测的工作原理是，假设已知数据集符合某种概率分布，然后用不一致检验来确定孤立点。不一致检验的应用需要事先知道数据集的参数（如正态分布），分布参数（如均值、标准差）及孤立点的数目。不一致性检验检查两个假设：工作假设和替代假设。工作假设是一个命题：n 个对象的整个数据集合来自一个初始的分布模型 F，H：$O_i \in F$，i=1，2，…… ，n，如果没有统计上显著证据支持拒绝这个假设，它就被保留。不一致性检验验证一个对象 O_i 关于分布 F 是否显著的大（或小）。根据不同数据所表征的信息不同，不同的统计量可以用做不一致检验。

- **基于偏离的方法**-基于偏离的孤立点检测的工作原理是，通过检查一组对象的主要特征来确定孤立点。给出的描述偏离的对象被认为是孤立点。一般采取两种探测技术：序列异常技术与 OLAP 数据立方体技术。

  序列异常技术是模仿人的思维模式，在观察一个连续序列后，迅速发现其中一项数据与其他数据的明显不同，即使不清楚数据的规则。已知 n 个对象的数据集(对象在数据库中为某条记录)，可以建立数据子集{s_1, s_2, ……, s_m}，$2 \le m < n$，由此来求得子集间的偏离程度，即“相异度”。由于事先不知道数据的总体特征，相异度函数的定义较为复杂。

  OLAP 数据立方体技术是在大规模的数据中采用数据立方体来确定反常区域。为了提高效率，偏离的探测过程与数据立方体的计算式重叠的。这个方法是发现驱动探索的一种方式，预先计算的指示数据异常的值被用来在集合计算的所以层次上指导用户进行数据分析。如果一个数据立方体单元值显著地不用于根据统计模型得到的期望值，该单元被认为是一个异常，并可采用可视化的提示来表示。用户可以选择对那些标为异常的单元进行钻取。一个单元的度量值可能反映了发生在立方体更低层次上的异常，这些异常从当前层次上是不可见的。而且，该模型隐藏了在数据立方体集合分组操作后面的异常情况。

- **基于距离的方法**-Knorr 和 Ng 给出具有一般意义的基于距离的孤立点定义和相应的孤立点挖掘方法,该方法不需要明确的数据分布,通过 k 邻居距离来确定是否孤立。比较典型的基于距离的孤立点定义有以下 3 个。

  定义 1(DB(*pct*, D)-Outlier) 如果数据集中至少有 *pct* 部分对象与对象 o 的距离大于 D,则对象 o 是一个基于距离的关于参数 *pct* 和 D 的孤立点,即 DB(*pct*, D)-Outlier。

  从定义可知,如果 o 在 D 范围内有不多于 N (1-*pct*)个邻居,则 o 是 DB(*pct*, D)-Outlier。用参数 D 确定对象 o 的邻域,参数 pct 判断对象 o 是否为孤立点。

  实际上，对于恰当定义的 *pct* 和 D，一个基于分布的孤立点定义同样可以利用 DB(*pct*，D)来定义,如定义 1 可以用 DB(0.9988,0.13σ)来表述；同时，它克服了基于统计的挖掘方法难以处理多维属性和要求用户预先知道数据集服从哪种统计分布模型的缺点。

  基于 DB(*pct*, D)的挖掘算法有基于索引算法、基于块嵌套循环算法和基于单元的算法。基于索引的算法采用多维索引结构(如 R 树或 k-d 树)来查找

每个对象 o 在半径 D 范围内的邻居。这个算法在最坏情况下的复杂度为 O(δN^2)，当维度 δ 增加时，复杂度的增加是线性的。但是，复杂度估算只考虑了搜索时间，而索引结构的构建是非常费时的。

基于块嵌套循环算法和基于索引的算法有相同的计算复杂度，但它避免了索引结构的构建,试图最小化 I/O 的次数，把内存的缓冲区分为两半,将数据集分为若干个逻辑块。通过精心选择逻辑块装入每个缓冲区域的顺序，可改善 I/O 效率。

基于单元格的 DB(*pct*, D)算法将 δ 维空间划分为边长为 D/(2√δ)的单元格，并以单元格为单位进行检测。其计算复杂度是 O(m(2√δ+1)$^{\delta}$+N)，其中 m 是单元个数,因此该算法仅适合于大数据集、低维度的场合。DB(*pct*, D)对参数 pct，D 比较敏感,而且缺少孤立程度的信息,因此难以度量和有效挖掘。

定义 2(top-n Outlier)如果一个数据集具有 N 个对象，给定对象孤立程度的计算公式，计算每个对象的孤立得分,孤立得分最高的 n 个对象就是所求孤立点,即 top-n Outlier。

在定义 1 和定义 2 基础上发展了以下两种定义。

定义 3 (top-n D^k-Outlier)数据集 O 中那些到其第 k 个最近邻居的距离 D^k 最大的 n 个对象就是孤立点,即 top-n D^k-Outlier。若 D^k(o)表示对象 o 与其第 k 个最近邻居的距离，则处于分布稀疏区域的数据点将具有较大的 D^k 值，而属于聚类中的类内数据点将具有较低的 D^k 值。D^k 孤立点挖掘方法基于各数据点 D^k 的排列,克服了 DB(*pct*，D)法缺少孤立程度信息的不足。同时，D^k 法无须用户指定距离参数 D。

在定义 3 的基础上还有一种基于划分的发现算法(Ramaswamy S，2000)。首先利用聚类算法划分数据集；然后计算各划分 P 的 D^k 边界(P.lower, P.upper)，使 P 中的每个点 *p*，满足 P.lower ≤ D^k(p) ≤ P.upper，并利用此信息确定 P 中是否可能包含孤立点；最后仅在可能包含孤立点的划分中计算和寻找孤立点。由于所要寻找的孤立点数目 n 相对较少，该方法可通过排除包含大量数据点的划分而降低计算量。实验已经证实，该方法关于 N 和 δ(≤10)的可扩展性均较好。但是，由于 D^k(p)并没有包含 p 点所有 k 个最近邻的全部信息,因而它并不能很好地反映其邻域的紧密或稀疏状况。

定义 4 (top-n w_k-Outlier)数据集 O 中那些与其 k 个最近邻居的距离之和 w_k 最大的 n 个对象就是孤立点,即 top-n w^k-Outlier。

对于数据点o,对象o 与其k个最近邻居的距离和称为o的权,记为$w_k(o)$。显然$w_k(o)$比$D^k(o)$更精确地度量了 o 的邻域的稀疏程度。

定义 3 和定义 4 利用排序，减少了距离参数 D 的输入，增加了参数 n。输出的孤立点的个数受 n 控制，但孤立点的顺序不受 n 影响，且易于确定。定义 3 仅考虑了第 k 个邻居的距离而忽略了最近邻居值，定义 4 考虑了所有邻居值,是基于最近邻居密度的计算,虽降低了计算速度,但提高了度量精度。

- **基于密度的方法-**上述孤立点定义是对数据集进行全局观察，孤立点挖掘方法均基于各数据点自身的邻域来判别其是否是孤立点，其检测标准是全局的、绝对的，因此所挖掘的孤立点是全局孤立点。但许多实际的数据集结构更复杂，还存在另一种孤立，这些孤立是相对于它们的局部邻域异常，因而被认为是“局部”孤立。图 1 是二维数据集，图中包含两个簇 C1，C2 和两个孤立点 o1，o2，其中 C1 稠密，C2 稀疏，o2 是全局孤立点，o1 是局部孤立点。根据上述定义及挖掘算法，o2 孤立点易于挖掘，但 o1 却难以挖掘，如果为了挖掘出 o1 而调整参数，那么 C1 中的大多数数据点都将被标识为孤立点。

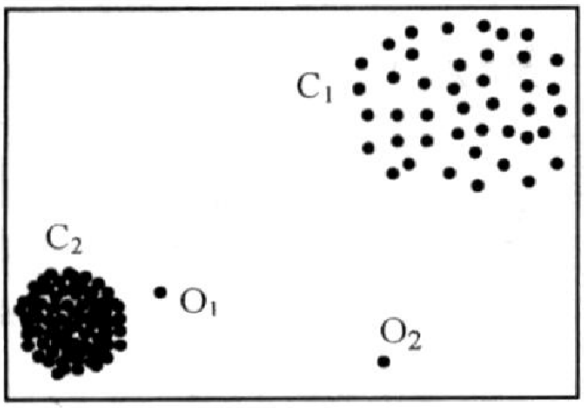

图 2 局部孤立示意图

为此，（Breunig et al，1999，2000）提出了局部孤立点概念和基于密度的孤立点定义,通过引入一个专门的度量单位: 孤立系数(OF: Outlier Factor),用局部孤立系数(LOF: Local Outlier Factor)来表征一个对象的局部孤立程度。在 LOF 算法中，根据给定的参数最少邻居数 k 和最近邻距离来确定邻域，通过计算对象的 k2 距离、可达距离和可达密度，用数据对象邻域的平均可达密度与其自身的可达密度之比表示 LOF，LOF 越大，其孤立程度越高。LOF 解决了局部孤立程度的度量和挖掘问题，同时摒弃了以往方法中数据对象非此即彼的概念。

（Jin et al，2006）提出了基于“反向 k 邻域”RNN_k(Reverse k Nearest Neighbors)的局部孤立度度量方法 INFLO(INFLuenced Outlierness)，不仅考虑数据点的 k 邻域，还考虑数据点的“反向 k 邻域”对数据孤立度的影响，从而避免数据分布复杂情况下 LOF 算法可能出现的错判。

为了克服 LOF 算法对于序列数据和低密度数据对象不能有效度量的缺陷，Tang 等提出基于连接的孤立系数(COF: Connectivity-based Outlier Factor)的方法，其算法是根据给定的参数最少邻居数 k 和数据对象的连接性来确定邻域，计算与其邻域的平均连接距离，用平均连接距离比作为基于连接的孤立系数 COF。虽可克服上述局限，但由于 COF 增加了连接路径的建立，因此计算比 LOF 更复杂。

LOF，INFLO 和 COF 等方法解决了局部孤立点定义与挖掘问题，但与基于距离的方法相比，其算法更加复杂，效率也较低，时间复杂度为 $O(kN^2)$ 的计算复杂度，其中 k 为邻居数，N 为数据点总数，难以用于大规模数据集；另外，检测结果对指定的参数 k-邻居的选择很敏感，当 k 值过小时，在孤立点彼此接近，形成一个小的孤立簇的情况下，会将这个小的孤立簇误判为正常数据簇，导致漏检；当 k 值太大时，接近稠密簇的孤立点可能会被误判为正常数据点，也会导致漏检。为了得到满意结果，需要反复调整参数 k，而每次调整参数均须重新构造邻域，邻域构造非常费时，具有 $O(kN^2)$的计算复杂度。为了改进算法的效率，降低对参数的敏感性，一些学者提出了避免距离或密度计算的方法、基于划分和剪枝技术的方法、基于属性划分的方法等等。

- 基于聚类的方法- 基于聚类的孤立点的定义是: 如果一个对象不属于任何簇，则该对象是基于聚类的孤立点。

聚类分析是数据挖掘的重要手段之一。目前有很多的聚类算法，如 DBSCAN，CLARANS，CHAMELEON，BIRCH，STING，Waver Cluster，CLIQUE 和 ROCK 等。簇的定义通常是孤立点的补，因此可同时发现簇和孤立点。但是它们的主要目标是产生聚类，即寻找性质相同或相近的记录并归为一个类。孤立点只是聚类分析的副产品而已。在实际应用中，可以先采用聚类算法对数据集进行聚类，然后再采用孤立点检测算法来发现孤立点。

表 1 对这几种类型的孤立点检测方法的主要特点进行了对比和分析。

表 1 几种孤立点检测算法的比较

检测方法	典型算法	特点(优缺点)	适用性
基于统计的方法	单样本多个离群检测算法(ESD)	易于理解,只对单维数据集，较为有效。需要知道数据集的先验知识。	不适合高维数据
基于偏离的方法	序列异常技术	概念有缺陷，遗漏了不少孤立点。	适用性不高
	OLAP 数据立方体技术	搜索空间大，人工探测困难。	可适用于多维数据
基于聚类的方法	聚类与孤立点检测算法结合	先聚类，后检测异常点。	适用于大规模的数据集
基于距离的方法	索引算法	I/O 代价较高，性能与索引结构有关	适用于低维空间的数据集
	NL 算法	理论上可以处理任意维，减少了算法的 I/O 次数。	适用于较低维 ,小规模的数据集
	单元算法	在低维空间时优于 NL 算法。	一般适用于较低维
基于密度的方法	LOF 算法	可以识别局部异常。	不适用于高维空间
	LSC 算法		

对于文献数据而言，已经有很多比较成熟地聚类分析方法，如共引聚类、耦合聚类和共词聚类等，因此基于聚类的方法在文献数据的孤立点挖掘中是比较有理论和实践基础的。

三、孤立点挖掘的过程

孤立点挖掘的过程可以粗略的分为四个步骤:

（一）、孤立点问题定义（Outlier Task Definition）：数据挖掘是为了在大量数据中发现有用的令人感兴趣的信息，因此发现何种知识就成为整个发现过程中第一个也是最重要的一个阶段。在问题定义过程中，数据挖掘人员必须和领域专家以及最终用户紧密协作，一方面明确实际工作对数据挖掘的要求；另一方面通过对各种学习算法的对比进而确定可用的算法。后续的学习算法选择和数据集准备都是在此基础上进行的。

（二）、数据收集和数据处理（Data Preparation and Preprocessing）：数据准备工作又可以分为三个子步骤：数据选取（Data Selection），数据 预处理（Data

Preprocessing）和数据变换（Data Transformation）。数据选取的目的是确定发现人为的操作对象，即目标数据（Target Data），是根据用户的需要从原始数据库中抽取的一组数据。数据预处理一般包括消除噪声、推导计算缺值数据、消除重复记录、完成数据类型转换（如把连续值数据转换为离散型数据，以便于符号归纳，或是把离散型转换为连续型等）。对于挖掘对象是数据仓库而言，数据交换的主要目的是数据降维（Dimension Reduction），即从初始特征中找出真正有用的特征，以减少数据挖掘时要考虑的特征和变量数。

（三）、孤立点挖掘：选择适合挖掘任务的算法。过程中应当考虑，（1）不同数据的不同特点；（2）满足用户和实际运行系统的要求。

（四）、解释并评估结果：阶段发现的模式可能存在冗余和无关的模式，需要根据相应的背景知识将其剔除；也有可能模式不能满足用户需求，则需要整个过程回退到前一阶段，如重新选取数据、采用新的数据变换方法、设定新的参数值，甚至要换新算法。

四、科学前沿的特征分析

Price 将在科学引文网络中频繁被引用的新近发表的文献称之为“研究前沿”（Research Front），并以此来描述学科研究领域的过渡本质。1994 年，瑞典科学计量学家 Persson 对研究前沿和知识基础作了界定：从文献计量学来看，引文（施引文献）形成了研究前沿，被引文献构成了知识基础。目前其他对“研究前沿”的定义，都是在 Price 和 Persson 的基础上进行的内涵和外延的扩展。陈超美将“研究前沿”定义为一组突现的动态概念和潜在的研究问题，并通过引文和共引的轨迹来追踪和识别研究前沿。然而，根据 Price 的定义，研究前沿需要由其在引文网络中的表现来判定，因此存在监测和识别的时滞。陈超美将这一问题进行了改进，将引文和关键词结合起来研究，既利用了引文分析的优势，也利用突发关键词所表达的新概念可以较早的监测到新研究前沿的出现。除此之外，研究前沿的出现往往伴随着其他文献计量学特征的出现，跟踪和探测这些文献计量学特征有可能揭示研究前沿的出现。

综合以上对“研究前沿”的定义，我们认为高被引论文和获诺贝尔奖的论文也是科学前沿的一种表现形式。希望通过对这两类数据的分析来研究科学前沿的属性特征。

4.1 高被引论文的分析

对数学、物理、化学、材料和空间五个领域高被引的论文的期刊和机构（第一作者机构）的分布进行了分析。

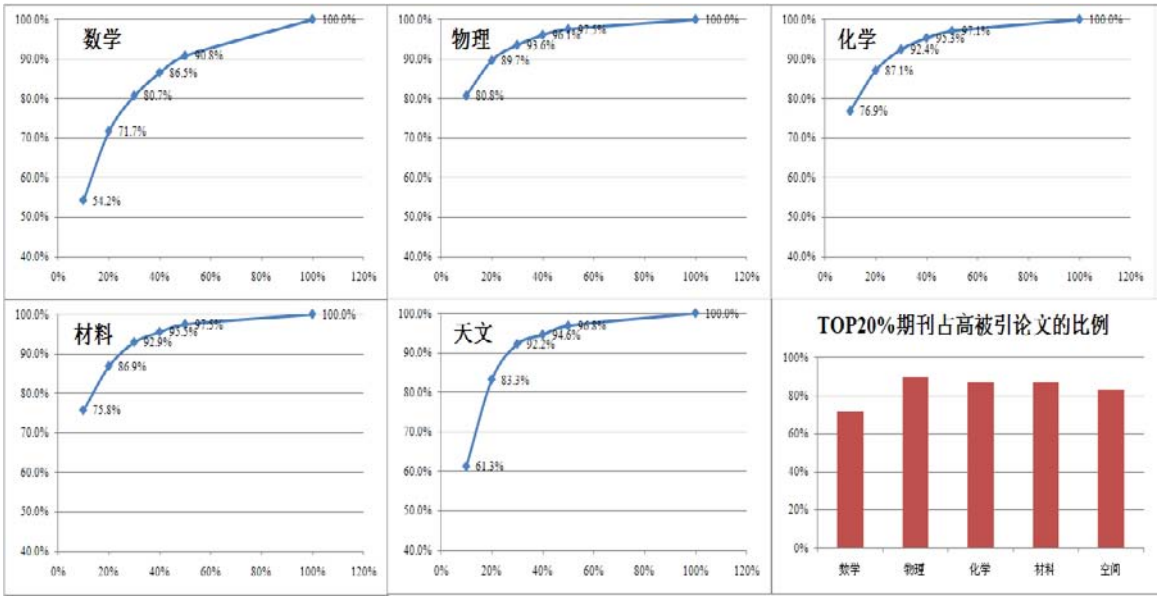

图 3 高被引论文的 TOP 期刊累积百分比

从五个学科高被引论文的期刊分布来看，其期刊集聚程度较高，其中 TOP10%的期刊对高被引论文的贡献都在 50%以上，其中数学最低，为 54.2%；物理最高 81.8%。相比较期刊而言，机构聚集程度稍低，其 TOP10%的机构对高被引论文的贡献在 40%以上，其中数学领域最低占 43%，化学领域最高占 61%。

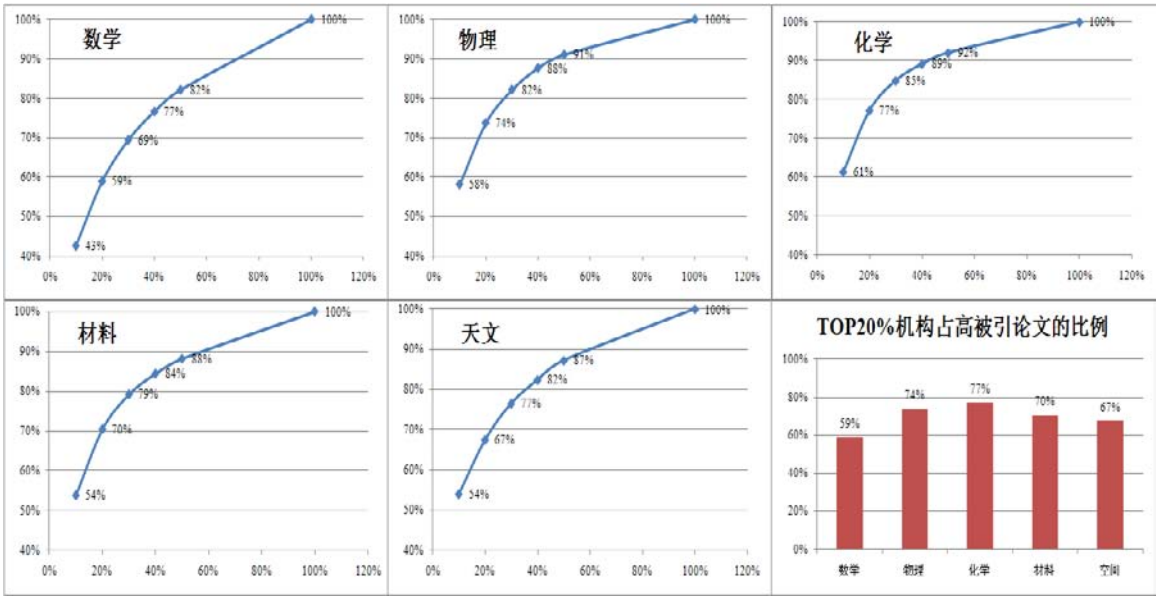

图 4 高被引论文的 TOP 机构累积百分比

4.2 诺贝尔奖论文分析

4.2.1 案例分析之 2010 年诺贝尔物理学奖论文

标题: Electric field effect in atomically thin carbon films

作者: Novoselov KS, Geim AK, Morozov SV, et al.

来源出版物: SCIENCE

卷: 306 期: 5296 页: 666-669

出版年:OCT 22 2004

被引频次: 3,386

参考文献时间特征:

分析其参考文献发表的时间特征，该文章引用参考文献在近 3 年发表（不包括发表当年，在统计数据时计算在内）的论文数量（11）占总参考文献数量（16）的 69%。

表 2 2010 年诺贝尔物理学奖参考文献相关信息

文章题名	期刊名称	期刊名称	主要作者	机构
Carbon nanostructures	2002	Critical reviews in solid state and materials sciences	Shenderova, oa	N Carolina State Univ

Carbon nanotubes - the route toward applications	2002	Science	Baughman, rh	Univ Texas
Cleavage of graphite to graphene	2001	Journal of materials science letters	Shioyama, h	AIST （National Institute of Advanced Industrial Science and Technology）
Electronic Transport-Properties Of Graphite, Carbons, And Related Materials	1981	Chemistry and physics of carbon	Spain, il	Univ maryland
Experimental evidence of a single nano-graphene	2001	Chemical physics letters	Affoune, am	Tokyo Inst Technol
Fabrication of mesoscopic devices from graphite microdisks	2001	Applied physics letters	Dujardin, e	Univ Bristol
Graphitic cones and the nucleation of curved carbon surfaces	1997	Nature	Krishnan, a	Nec Res Inst（日本 NEC 美国研究中心）
Hall constant in quantum-sized semimetal Bi films: Electric field effect influence	2000	Journal of applied physics	Butenko, av	Bar Ilan Univ
Intercalation compounds of graphite	2002	Advances in physics	Dresselhaus, ms	
Molecular electronics: From devices and interconnect to circuits and architecture	2003	Proceedings of the ieee	Stan, mr	Univ Virginia
Nanotube molecular wires as chemical sensors	2000	Science	Kong, j	Stanford Univ
Novel electronic wave interference patterns in nanographene sheets	2002	Journal of physics	Harigaya, k	AIST
Organic thin-film transistors: A review of recent advances	2001	Ibm journal of research and development	Dimitrakopoulos, cd	IBM Corp
Possibility of a metallic field-effect transistor	2004	Applied physics letters	Rotkin, sv	Univ Illinois
STM Investigation Of Single Layer Graphite Structures Produced On Pt(111) By Hydrocarbon Decomposition	1992	Surface science	Land, ta	Forschungszentrum Julich
Sensitivity of single multiwalled carbon nanotubes to the	2003	New journal of physics	Kruger, m	Univ Zurich

environment				

4.2.2 案例分析之 2009 年诺贝尔物理学奖论文

标题: DIELECTRIC-FIBRE SURFACE WAVEGUIDES FOR OPTICAL FREQUENCIES

作者: KAO KC, HOCKHAM GA

来源出版物: PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS-LONDON

卷: 113 期: 7 页: 1151-&

出版年: 1966

被引频次: 205

该文章引用近 3 年发表的论文 5 篇，占 50%。

表 3 2009 年诺贝尔物理学奖参考文献相关信息

文章标题	年	期刊	主要作者	机构
Single-conductor surface-wave transmission lines	1951	Proceedings of the institute of radio engineers	Goubau, G.	Signal Corps Engineering Laboratories
An experimental investigation of the properties of corrugated cylindrical surface waveguides	1954	Proceedings of the institution of electrical engineers-London	Barlow, H. E. M., and Karbowiak, A. E.	Electrical Engineering at University College，London
Azimuthal surface waves on circular cylinders	1955	Journal of applied physics	Elliott, R. S.	
Light scattering by glasses	1956	J chem phys	Maurer, R. D.	
Field theory of guided waves	1960	McGraw-Hill	. Collin, R. E.	
An investigation of higher order surface waves on cylindrical structures'	1961	Ph.D. Thesis	Savard, J. Y.	University College, London
Propagation in the azimuth direction of a cylindrical surface wave	1963	Ph.D. Thesis	Potter, S. V.	University College London
Some observations on the absorption of iron in	1965	Phys. Chem. Glasses	Steele, F. N., and Douglas,	

silicate and borate glasses			R. W.	
The diffraction of a cylindrical surface wave by surface discontinuities	1965	Ph.D. Thesis	Breithaupt, R. W.	University College London
Coupling of optical fibres and scattering in fibres	1965	Opt. Soc. Amer.	Jones, A. L.	
Characteristics of modulated cylindrical surface waveguide		Not published until 1966	Hockham, C	

表 4 近年来诺奖论文的部分指标分析

获奖时间	发表时间	参考文献的新颖程度（引用近 3 年发表文献的比例）	TOP10%机构所占的比例	TOP20%期刊所占比例
2010	2004	69%	60%	56%
2009	1966	60%	——	10%
2008	1961（两篇）	88.60%	——	33%
2007	1988、1989（两篇）	66.70%	——	71%
2006	1992	58.60%	——	32%（其中 14 篇来自 ASTROPHYSICAL JOURNAL）

诺贝尔奖可以很大程度上反应论文研究结果的前沿性。这里对其参考文献的一些计量特征进行了分析（见表 4），利用参考文献中近 3 年论文的比例来度量论文对最新研究成果的吸收和继承情况;参考文献中TOP10%[2]机构发表的比例来度量其对重点机构重要研究成果的跟踪能力,参考文献中TOP20%[3]期刊刊登的比例来度量其对重要期刊重要研究成果的跟踪能力。

统计结果表明，诺贝尔奖论文的参考文献较新颖，新颖程度平均可以达到 69%。

由于有些诺贝尔物理奖论文发表年代比较早,其参考文献中的机构多数不全,因此就没有做深入分析，但是以 2010 年诺贝尔物理学奖的论文为例，其参考文献中 TOP10%机构发表的论文占到 60%。这里我们选择的机构是从整个物理领域出发，如果聚集到具体的领域方向，结合专家的判断，这个比例将会更高。

[2] TOP10%根据 Web of knowledge 中 ESI 数据库物理领域中发表高被引论文前 10%的机构。
[3] TOP20%根据 Web of knowledge 中 ESI 数据库物理领域中刊登高被引论文前 20%的期刊。

TOP20%期刊所占比例的分析结果表明，总体来说参考文献中 TOP20%期刊发表的文章占有较高的比例。其中比例较低的情况，如 2009 年诺贝尔物理学奖的论文虽然TOP20%期刊的比例只有10%,然而其参考文献中有3篇是博士论文，有一篇当时没有发表;如2006年诺贝尔物理学奖论文TOP20%期刊的比例为32%，但是其中 14 篇（48.3%）来自 ASTROPHYSICAL JOURNAL，该刊虽然没有进入物理学高被引期刊 TOP20%列表，但是他是天体物理学领域世界顶级的期刊。同样，如果期刊选择再聚集一些，这个比例将会很高。

五、案例研究

本研究选取钍基核燃料循环的数据进行文献数据孤立点的分析。

5.1 基于统计的孤立点定义

利用钍基核燃料循环的数据[4]进行基于统计的孤立点定义。考虑到作者关键词是作者认为其研究成果中最有创新性，需要大家特别注意的部分内容，因此将作者给出的关键词（1214）作为分析对象进行分析、测试。其中，作者关键词数量小于文献数量的原因是文献检索为全时间范围，较早的文献中部分关键词缺失，但从研究具有持续性的角度，由此带来的关键词的分析误差可以忽略不计。

这里我们定义钍基核燃料循环数据的分布模型 F，假设关键词构成数据集合 U:

$$Oi \in F，（i=1,2，……，n，and\ Oi \notin U）$$

[4]本次分析采用的数据库为 Web of Science（包括 Science Citation Index Expanded 和 Conference Proceedings Citation Index 数据库），利用关键词结合领域分类的方式检索了所有在钍基核燃料循环研究方面发表的论文。数据检索时间为 2010 年 11 月 22 日，共检索到有效数据 1422 条。

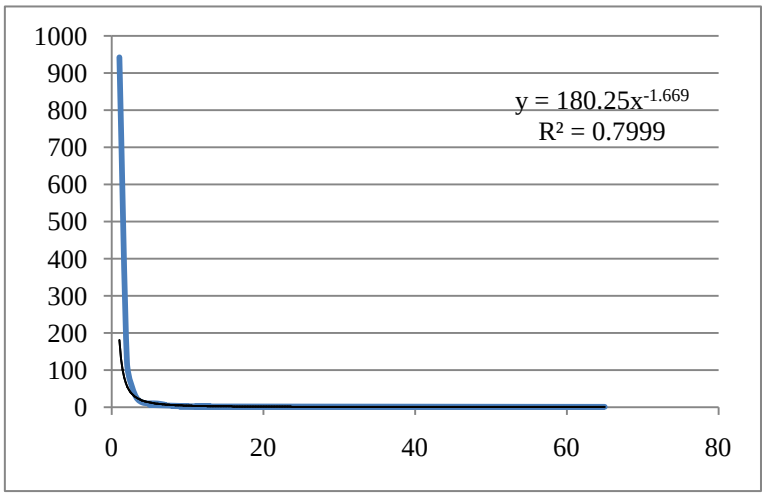

图 5 钍基核燃料循环关键词词频分析

首先，我们对钍基核燃料循环关键词的词频分布进行了分析（见图 5），结果显示关键词的词频呈幂分布，即高频词占有很小的比例，而低频词占绝大多数。如果从关键词含义的角度来理解，低频词应当是我们关注的孤立信息，它们在整个词集中表现为独特的信息，而从频率分布的角度来看，低频词是不具有孤立信息的，因此并不适合利用直方图来检测孤立点。因此需要对关键词进行一定的规范和处理并且对孤立点的定义进行范围限定才能得到我们想要的结果。

这里根据关键词的频次特征，定义最新出现的低频关键词作为用于分析前沿特征的孤立点信息。

孤立点选取的步骤为:

1、对关键词进行清洗，合并同义词等。

2、找出所有在关键词集只出现一次的关键词。

3、选择最近一年（2010 年）出现一次的关键词。

分析结果表明，2010 年发表的文章中新出现的关键词有 72 个（见表 3），分布在 30 篇文章中（见表 4），一般情况下一篇文章中包含多个新关键词。

表 3 2010 年发表的文章中出现的新关键词

Spectrophotometry	Chemical recoveries
Tissue Paper matrix	Extractant-impregnated resins
TODGA	Extraction chomatography
Dissolver solution	Amberlite IRA-410
L-arginine	Nuclear reactor materials
Tributylphosphate	Amberlite XAD-2
TRISO fuel	Olive cake
Lead recovery	Oxidation state

TTA	Column-mode separation
Lead-208	Peaceful nuclear explosive
Supercritical CO_2	Phase diagrams
Supported liquid membrane	Analytical waste
Dynamic ion-exchange chromatography	Conditional interaction constant (beta)
Liquid extraction	Plutonium purification
LWR spent fuel	Poly[dibenzo-18-crown-6]
2-(5-bromo-2-pyridylazo-5-diethylaminophenol	FTIR
East reactors	Pre-equilibrium reactions
Actinide equilibrium	Precipitation
Actinide evolution	Fuel Rod Performance Code
Modified polymers	Full exciton model
THOREX	Arsenazo III
Ce(III)	GDH model
Eosin B	Green procedure
Natural radionuclides	H_2O_2
2-Ethylhexyl-2-ethylhexyl phosphonic acid	Automated separation system
Natural waters	Cyanex272
Equilibrium reactions	Semi-empirical calculation
Ceramic nuclear fuel	Sequential extraction
Neutron emission cross section	Simulated waste
Chelex-100	Sintered (Th)
Non-proliferation	Hollow fibre
TBEP	HPLC
TBP-HNO_3 adduct	Solvation mechanism
Deuteron-induced reactions	Humic acid
DHOA	Spectrophotometry
Thermodynamic modeling	ICP-OES

表 4 包含新关键词的文章标题

标题
Actinide evolution and equilibrium in fast thorium reactors
Alpha spectroscopic analysis of actinides (Th, U and Pu) after separation from aqueous solutions by cation-exchange and liquid extraction
Comparative extraction efficiencies of tri-n-butyl phosphate and N,N-dihexyloctanamide for uranium recovery using supercritical CO_2
Comparison of two sequential extraction protocols for fractionation of natural radionuclides in soil samples
Considerations on safety and proliferation-resistant aspects for the MSBR design
Cyanex272 impregnated on Amberlite XAD-2 for separation and preconcentration of U(VI) from uranmicrolite (leachates) ore tailings

Determination of (n,2n) Reaction Cross Sections for Some Nuclei with Asymmetry Parameter
Determination of alkali, alkaline earth and transition metal ions in UO_2, ThO_2 powders and sintered (Th,U)O_{-2} pellets by ion chromatography
Development of a relatively cheap and simple automated separation system for a routine separation procedure based on extraction chromatography(Olive cake)
Effect of ionic strength on complexation of Pu(IV) with humic acid
Effect of Using Thorium Molten Salts on the Neutronic Performance of PACER
Excitation Functions of Some Neutron Production Targets on (d,2n) Reactions
Exploring new coolants for nuclear breeder reactors
Extraction of U(VI) and Th(IV) from nitric acid medium using tri(butoxyethyl) phosphate (TBEP) in n-paraffin
Extraction studies of plutonium from acidic solution using gamma-ray induced PC-88A/TBP modified polymers
Preconcentration of uranium(VI) and thorium(IV) from aqueous solutions using low-cost abundantly available sorbent
Rapid separation of lanthanides and actinides on small particle based reverse phase supports
Selective extraction of Th(IV) over U(VI) and other co-existing ions using eosin B-impregnated Amberlite IRA-410 resin beads
Selective ion exchange separation of uranium from concomitant impurities in uranium materials and subsequent determination of the impurities by ICP-OES
Selective recovery of uranium from THOREX feed by hollow fibre supported liquid membrane technique containing di(2-ethylhexyl) isobutyramide (D2EHIBA) as the carrier
Separation and purification of americium from analytical waste solutions
Separation of uranium from different uranium oxide matrices employing supercritical carbon dioxide extraction （TTA， Tissue Paper matrix）
Simplified alpha-spectroscopic analysis of uranium in natural waters after its separation by cation-exchange
Sorption study of U(VI), Th(IV) and Ce(III) on poly[dibenzo-18-crown-6] in l-arginine to develop sequential column chromatographic separation method
Spectrophotometric determination of trace amount of uranium (VI) in different aqueous and organic streams of nuclear fuel processing using 2-(5-bromo-2-pyridylazo-5-diethylaminophenol
Studies on purification of plutonium from silver ions
Study on radiogenic lead recovery from residues in thorium facilities using ion exchange and electrochemical process（lead 208，lead recovery）
Thermodynamic assessments of the Al-Th and Th-Zn systems
Under irradiation issues of the CSZ-based inert matrix fuels from IFA-652 Halden experiment
Utilization of TRISO Fuel with LWR Spent Fuel in Fusion-Fission Hybrid Reactor System

The optimized procedure was extended to remove uranium from simulated tissue paper waste matrix smeared with uranium oxide solids.

这里对所有 30 篇孤立点数据进行分析，从关键词的新颖性来判断，近期研发活跃的国家是印度、土耳其和塞浦路斯这三个国家。而我们通常所知道的核燃料循环研究实力较强的美国、法国和日本却没有最新的研究产生，可能也说明钍基核燃料循环并非是一个非常大众的技术选项，或者发达国家有更好的核能或能源发展选择。

表 5 初选孤立点数据的国家分布

排序	记录数	第一国家
1	12	INDIA
2	5	TURKEY
3	2	CYPRUS
4	1	BELGIUM
5	1	BRAZIL
6	1	ENGLAND
7	1	GERMANY
8	1	HUNGARY
9	1	IRAN
10	1	ITALY
11	1	PAKISTAN
12	1	PEOPLES R CHINA
13	1	SLOVENIA
14	1	SPAIN

这些关键词请相关的专业人员进行判读，从其专业角度给出他们认为比较有意思的一些关键词（表 3 中标红的部分内容），在这些文章中找出了和专业人员给出关键词对应的文章（表 4 中标红的部分内容），并进行了详细解读。

有意义的孤立点的概念分析:

1、Development of a relatively cheap and simple automated separation system for a routine separation procedure based on extraction chromatography(Olive cake)

该研究利用榨橄榄油后的油渣饼进行钍乏燃料的提取，此过程的原理和方法更加倾向于是一个物理过程，与其他萃取剂的提取是基于不同的知识基础，另外它可以用废料来处理废料，可能成为未来提取钍乏燃料一个有前景的方法。

2、Study on radiogenic lead recovery from residues in thorium facilities using ion exchange and electrochemical process（lead 208，lead recovery）

钍燃料作为反应堆的燃料需要一个纯化过程，纯化过程将产生大量的钍残留物，利用离子交换和电化学过程可以从这些残留物中提取大量放射性的铅-208，论文中提到的方法可以提取硝酸介质中熔解的 98%的铅。放射性铅-208 在生产新元素方面具有重要作用，有可能成为战略材料。

3、Separation of uranium from different uranium oxide matrices employing supercritical carbon dioxide extraction （TTA， Tissue Paper matrix）

本研究利用超临界二氧化碳和 TBP-HNO（3）的混合物从不同的铀二氧化物中提取铀，不仅如此，还能从被放射污染的废纸中提取出固体氧化铀。该方法向甲烷中加入了 2.5%的 TTA 后，几乎能提取出铀氧化物中全部的铀。

4、Under irradiation issues of the CSZ-based inert matrix fuels from IFA-652 Halden experiment（Fuel Rod Performance Code）

本研究中 ENEA 利用重沸水堆上的 IFA-652 Halden 实验对不同的惰性基质燃料，如钙稳定氧化锆和氧化钍进行了测试，研究结果证明在功率曲线的实验数据集基础上，超铀燃料的性能编码与实验结果一致。

5.2 基于聚类的孤立点定义

由于最新的发表文章还没有引文聚集的特征，而且分析其参考文献也能从关系的角度对文献进行类别的划分，因此这里我们利用其耦合关系进行聚类，从而分析数据的孤立点信息，并之前的分析结果进行对比。

图 6 是利用这 30 篇文章的耦合关系[5]进行聚类而得到的关系图，可以看出不同关系定义得到的孤立点的信息是不同的。然而，上一节得到的 1、2 和 4 三篇篇文章中的关键词仍被作为孤立点识别出来，主要原因是其研究基础或方法原理与其他研究有差异。同样也说明，这样的研究结果在我们的数据集中出现具有现实意义。

[5] 指利用参考文献的

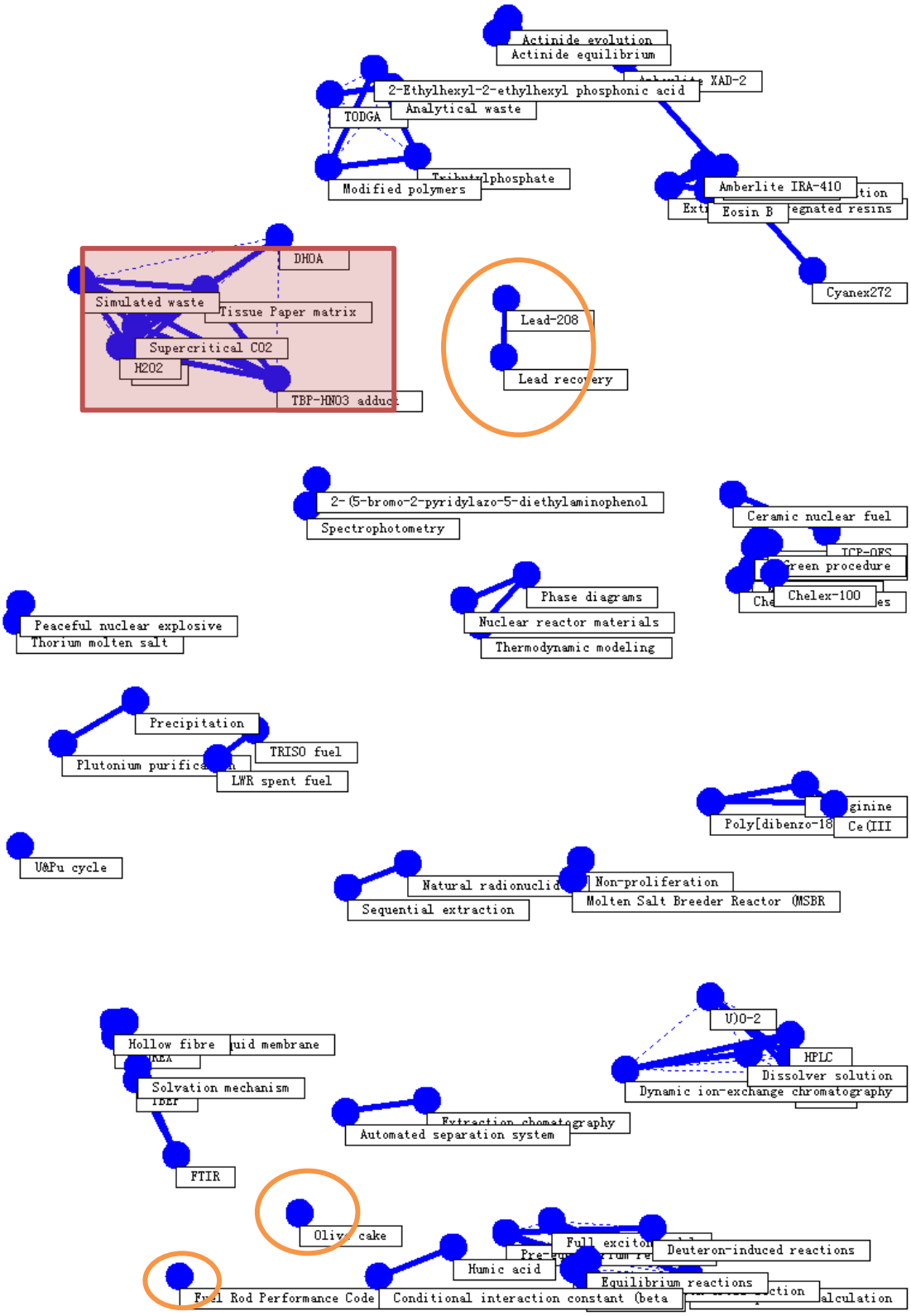

图 6 基于耦合聚类的孤立点分析结果

5.3 基于时间属性的孤立点定义

根据论文数据的独有特征[6]，本研究提出了基于时间属性的孤立点定义。将某一领域最新（根据分析时间窗来定）发表的论文都定义为孤立点。并利用全部文献的聚合信息，来定义较有影响力的期刊和机构，并利用相关指标对有意义的孤立点数据进行检测和筛选。

利用检索到的 1422 条有效数据，对这些数据的期刊和机构聚合特征进行了统计分析。根据期刊刊载和机构产出的特征，我们定义聚合特征明显的目标对象为具有较大影响力的对象，在一定程度上对于数据的选取具有参考意义。

对钍基核燃料循环的论文进行了分析，见表 6 所示，TOP10 的期刊刊载了所有论文的 68%，表明这些期刊已经得到了钍基核燃料循环相关研究人员的认可，他们乐于将自己的研究成果发表在这些期刊上。表 7 给出了钍基核燃料循环产出 TOP20 的机构，这些机构共发表了 446 篇文章，占所有论文的 31%，从这些机构的分布可知，它们都是在该领域研究中表现突出的研究机构。

首先对 1422 篇论文进行时间属性的筛选，最新的一年有 57 篇论文发表，我们定义它们为初选孤立点集，然后根据机构和期刊属性对初选集进一步选择，最终选取有意义的孤立点。

表 6 钍基核燃料循环论文的期刊分布（TOP10）

序号	论文数	期刊名（缩写）	累计百分比
1	173	*Journal of Radioanalytical and Nuclear Chemistry*	12.17
2	142	*Journal of Nuclear Materials*	22.15
3	130	*Radiochimica Acta*	31.29
4	119	*Transactions of the American Nuclear Society*	39.66
5	83	*Nuclear Technology*	45.49
6	81	*Annals of Nuclear Energy*	51.20
7	72	*Journal of Radioanalytical and Nuclear Chemistry-Articles*	56.26
8	66	*Nuclear Science and Engineering*	60.9
9	60	*Physical Review C*	65.12
10	44	*Journal of Nuclear Science and Technology*	68.21

[6]作者注：论文是某一学术课题在试验性、理论性或预测性上具有新的科研成果或创新见解和知识的科学记录，或是某种已知原理应用于实际上取得新进展的科学总结，用以提供学术会议上宣读、交流、讨论或学术刊物上发表，或用作其他用途的书面文件。发表的论文通常都承载了一个最新研究成果，都是科学家经过严谨的实验和深入的思考后对自己研究工作形而上的总结，而且研究成果经过同行评议。

表 7 发表论文数量前 20 的机构的文献计量特征

序号	机构名	所属国家	论文数	总被引	篇均被引	H 指数
1	Bhabha Atom Res Ctr	India	98	406	4.1	10
2	Japan Atom Energy Res Inst	Japan	36	317	8.8	11
3	Gazi Univ	Turkey	34	293	8.6	9
4	Oak Ridge Natl Lab	USA	31	376	12.1	5
5	Atomic Energy Commission	France	29	208	7.2	9
6	CNRS	France	28	229	8.2	10
7	Los Alamos Natl Lab	USA	24	467	19.4	9
8	Forschungszentrum Julich	Germany	22	59	2.7	4
8	Indira Gandhi Ctr Atom Res	India	22	106	4.8	6
8	Joint Inst Nucl Res	Russia	22	307	14.0	6
11	Kyoto Univ	Japan	21	130	6.2	6
12	Tokyo Inst Technol	Japan	19	50	2.6	4
13	Univ Paris 11	France	19	300	15.8	10
14	Argonne Natl Lab	USA	18	433	24.1	4
15	Commiss European Communities	European	18	162	9	7
16	Forschungszentrum Karlsruhe	Germany	17	218	12.8	7
16	Atom Energy Canada Ltd	Canada	17	313	18.4	4
16	Lawrence Berkeley Natl Lab	USA	17	71	4.2	5
16	Lawrence Livermore Natl Lab	USA	17	542	31.9	7
20	Atom Energy Author	Egypt	16	41	2.6	4

这里，根据机构（见表 7 机构名属性）对这 57 篇论文进行选择，筛选出 20 篇论文。再根据期刊属性对这 20 篇论文的数据集进一步选择，得到 14 篇论文。

表 8 利用时间、机构和期刊属性筛选得到的孤立点论文集

Comparative extraction efficiencies of tri-n-butyl phosphate and N,N-dihexyloctanamide for uranium recovery using supercritical CO_2
Determination of alkali, alkaline earth and transition metal ions in UO_2, ThO_2 powders and sintered (Th,U)O-2 pellets by ion chromatography
Development of a relatively cheap and simple automated separation system for a routine separation procedure based on extraction chromatography
Extraction of U(VI) and Th(IV) from nitric acid medium using tri(butoxyethyl) phosphate (TBEP) in n-paraffin
Extraction studies of plutonium from acidic solution using gamma-ray induced PC-88A/TBP modified polymers
Nanosecond laser ablation of Thoria fuel pellets for microstructural study
Rapid separation of lanthanides and actinides on small particle based reverse phase supports
Selective recovery of uranium from THOREX feed by hollow fibre supported liquid membrane technique containing di(2-ethylhexyl) isobutyramide (D2EHIBA) as the carrier
Separation and purification of americium from analytical waste solutions
Separation of uranium from different uranium oxide matrices employing supercritical carbon dioxide extraction
Spectrophotometric determination of trace amount of uranium (VI) in different aqueous and

organic streams of nuclear fuel processing using 2-(5-bromo-2-pyridylazo-5-diethylaminophenol
Studies on purification of plutonium from silver ions
Thermodynamic and transport properties of thoria-urania fuel of Advanced Heavy Water Reactor
Verification of the surrogate ratio method

与之前的筛选结果进行比较，最终的选取结果与前两次的分析均有部分重合，结果集要大于前两次，因此还需要定义筛选指标进一步缩小目标结果集。

六、结果和建议

有意义的文献孤立点的查找可以使目前科学前沿信息监测的内容更加全面。然而，不同孤立点定义得到的信息结果不同，对于孤立点检测的定义及如何探测有意义的孤立点还需要进行深入的研究和分析。本研究对高被引论文的期刊和机构信息进行了统计分析，研究表明针对某一特定领域，期刊和机构属性能在很大程度上反映研究成果的质量和影响力。并对 2006-2010 年诺贝尔物理学奖论文的参考文献的相关指标进行了分析，分析结果也表明参考文献的时间、期刊和机构属性也能从一定程度上体现论文的质量和影响力。

本研究利用基于统计、基于聚类和基于时间属性的孤立点定义对文献数据进行了分析，分析结果表明基于统计的孤立点分析以新关键词作为分析对象，可以得到学科领域最新研究进展的信息；基于耦合聚类的孤立点分析将参考文献的关系作为聚类的主要指标，因此可以从找出与其他文章的研究基础和方法原理有明显差异的文献数据；而基于时间属性的孤立点定义，由于扩大了孤立点的筛选范围，因此要进一步细化指标定义。

另外，在对有意义的孤立点进行筛选时，本研究仅使用了单指标或多个单指标的综合判断，没有利用复合指标。但是如果要实现自动化检测，复合指标的定义是必要的，这样可以相对简单，并且准确快速的实现孤立点的检测。

参考文献

Vasiliki Karioti, Doctoral candidate.The Analysis of Outliers in Statistical Data. http://www.ntua.gr/eseve/Vasikh_Ereyna/Thalis/Thalis_projects_English_summaries/Karoni.pdf

夏勇. 聚类分析和孤立点识别技术研究及其应用.黑龙江: 哈尔滨工业大学，2008.

薛安荣，姚林，鞠时光，陈伟鹤，马汉达. 孤立点挖掘方法综述，计算机科学, 2008, 35(11):13-18.

Ramaswamy S , Rastogi R , Kyuseok S. Efficient algorit hms for mining outliers f rom large data set s //Proc. of the ACM SIG2MOD International Conference on Managem ent of Data. Dallas, 2000: 4272438

Breunig M M, Kriegel H P , Ng R T , et al . OPTICS-OF : identifying local outliers //Proc. of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases. Lecture Notes in Computer Science 1704, Prague, 1999: 2622270

Breunig M, Kriegel H P, Ng R ,et al . LOF : Identifying densitybased local outliers //Proc. of ACM SIGMOD Conference. Dallas, 2000: 932104

Price, D.J., Networks of scientific papers. Science, 1965. 149(3683): 510-515.

Persson, O., The intellectual base and research fronts of JASIS 1986 - 1990. Journal of the American Society for Information Science and Technology, 1994. 45(1): 31-38.

Chen, C. Measuring the Movement of a Research Paradigm. in Proceedings of SPIE-IS&T: Visualization and Data Analysis 2005. 2005. San Jose: SPIE- The International Society for Optical Engineering.

Chen, C., CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. Journal of the American Society for Information Science and Technology, 2006. 57(3): p. 359-377.

陈仕吉, 科学研究前沿探测方法综述. 现代图书情报技术, 2009. 9: p. 28-33.